Characterization of low thermal budget silicon MOSFETs for Digital and High frequency application on 3D sequential integration systems

Tadeu Mota Frutuoso

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : EEATS - Electronique, Electrotechnique, Automatique, Traitement du Signal (EEATS)
Spécialité : Nano électronique et Nano technologies
Unité de recherche : Laboratoire d'Electronique et de Technologie de l'Information (LETI)

Caractérisation des mosfets en silicium à budget thermique réduit pour applications numériques et haute fréquence sur des systèmes d'intégration séquentielle 3D

CHARACTERIZATION OF LOW THERMAL BUDGET SILICON MOSFETS FOR DIGITAL AND HIGH FREQUENCY APPLICATION ON 3D SEQUENTIAL INTEGRATION SYSTEMS

Présentée par :

Tadeu MOTA FRUTUOSO

Direction de thèse :

Philippe FERRARI Professeur des universités, Université Grenoble Alpes	Directeur de thèse
José LUGO ALVAREZ CEA-Leti	Co-encadrant de thèse
Xavier GARROS Ingénieur HDR CEA-E5, CEA	Co-encadrant de thèse

Rapporteurs :

Jean-Pierre RASKIN
PROFESSEUR, Université Catholique de Louvain
Jean-Michel SALLESE
SENIOR SCIENTIST, Ecole Polytechnique Fédérale de Lausanne

Thèse soutenue publiquement le **27 septembre 2023**, devant le jury composé de :

Philippe FERRARI PROFESSEUR DES UNIVERSITES, Université Grenoble Alpes	Directeur de thèse
Florence PODEVIN PROFESSEURE DES UNIVERSITES, Université Grenoble Alpes	Présidente
Jean-Pierre RASKIN PROFESSEUR, Université Catholique de Louvain	Rapporteur
Jean-Michel SALLESE SENIOR SCIENTIST, Ecole Polytechnique Fédérale de Lausanne	Rapporteur

Invités :

Xavier Garros
directeur de recherche CEA, CEA-LETI
Chevalier Pascal
INGENIEURE DOCTEUR, STMicroelectronics

CHARACTERIZATION OF LOW THERMAL BUDGET SILICON MOSFETS FOR DIGITAL AND HIGH FREQUENCY APPLICATION ON 3D SEQUENTIAL INTEGRATION SYSTEMS

Tadeu Mota Frutuoso

Xavier GARROS

Jose LUGO

Philippe FERRARI

University of Grenoble Alpes

2023

Abstract

3D sequential integration (3DSI) consists of sequentially stacking active device layers using vertical interconnections with similar dimensions as standard Back-End-Of-Line contacts (<100nm). It allows the co-integration of different systems on separated layers with a high interconnection density and it eliminates costly trade-offs coming from the optimization of different devices on the same substrate. Likewise, the reduced interconnection parasitic and heterogeneous integration offer great potential for 5G millimeter-wave (mmW) applications.

However, 3D stacked devices come along with new process challenges. Top-tier transistors need to be processed at low temperatures (≤ 500°C) to preserve the integrity of devices on lower tiers. Standard CMOS integration with low thermal budget (LTB) leads to substantial performance degradation. Nevertheless, new breakthroughs in the silicon LTB integration process open the path to the development of devices that reach the same performance of their high temperature counterparts. Therefore, the objective of this Ph.D. work is to analyze the effects of those new processes on the electrical characteristics of LTB MOSFET devices and draw guidelines for further optimization.

The manuscript for this Ph.D. introduces the main results obtained from the recent development of this technology and it is presented on three parts:

Activation of source and drain dopants near the junction using a low temperature Solid State Epitaxy Recrystallization (SPER) anneal. The study is performed with an estimation of the junction profile using a novel nondestructive CV technique coupled to as improved conformal mapping model of the transistors fringe capacitances. The results are used to understand the electrical behavior and degradation mechanisms of the devices as function of the overlap position.

Trapping properties of the low permittivity material (SiCO) used for the low temperature gate spacer oxide and its effects on the transistor performance. Two trapping mechanisms are identified on this material: Fast Silicon interface traps, related to the quality of the native oxide, and slow deep defects distributed in the bulk of the SiCO oxide. The effect of those traps near the access region of the electrical performance transistor are studied.

Effect of key low temperature process steps of the devices RF FoMs. The objective is to evaluate the performance of the devices at high frequencies. The lower parasitic capacitances from SiCO spacers, low gate resistance from the UV nanoseconds laser anneal and high mobility from the CESL tensile stain are key low thermal budget steps contributing to high-performance RF transistors with similar FoMs to HTB counterparts.

Table des matières

Statement of Original Authorship

The work contained in this thesis has not been previously submitted to meet requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signature: _____

Date: _____

Chapter 1: Introduction

1.1 3D SEQUENTIAL INTEGRATION

1.1.1 Context for 3D integration

As the scaling of logic nodes continues, each new generation must add more functions in order to maintain profit margins by reducing assembly costs. However, as the number of CMOS-compatible functions available decreases, traditional scaling methods are expected to reach their limits. So far, the number of devices per unit footprint is expected to increase by the end of the decade. This is primary due to reductions in gate and metal pitches as well as the introduction of a third dimension. However, this scaling route may become increasingly challenging as the gate pitch, gate length, and the number of stacked devices reach 40nm, 12nm and 4 respectively. This is due to the increase of process complexity, which also come along yield reduction and in turn increased cost. [IRDS and A. Mallik].

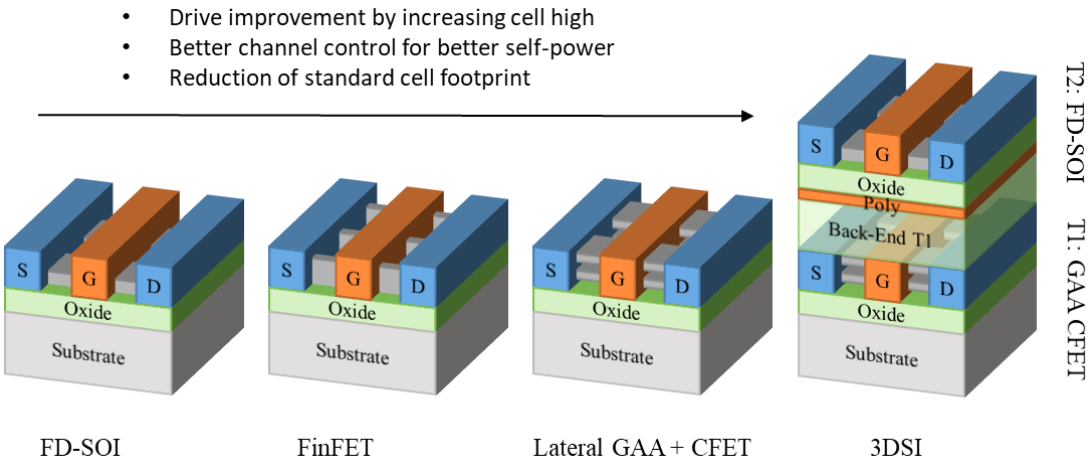


Figure 1-1 Sketch illustrating the evolution the device architecture of digital nodes. As cost and area reductions diminish, 3D sequential integration is expected to bring cost-advantage by adding SoCs complementary functions on top tiers that are integrated using non-scaled technologies.

With the saturation of the pitch reduction, there will be no room for further 2D geometry scaling of a System on a Chip (SoC). To overcome these limitations, the development of 3D integration process such as device-over-device stacking, fine-pitch layer transfer, or monolithic 3D is necessary. These strategies will not only maintain system performance and power gains, but also potentially preserve cost advantages by

integrating non-scaled components on different layers. Furthermore, it allows the use of the most appropriate technology (high mobility Ge and III-V, 2D materials) for the heterogeneous integration of each layer of the system. As predicted by the International Roadmap for devices and Systems, the transition to 3D integration of complementary System-on-Chip (SoC) functions is expected to take place by 2031, after the deployment of CFET GAA 1.5nm nodes. The transition from 2D to 3DVSLI is projected to evolve trough the following generations: (1) Logic 3D SRAM or MRAM stack; (2) Analog, RF and I/O stack (3) True-3D VLSI: Clustered functional stacks.

Increasing the memory density in a device through the integration of logic 3D SRAM (Static Random Access Memory) and MRAM (Magnetic Random Access Memory) stack is a promising strategy for 3D integration. By stacking memory layers on top of a logic layer in a 3D configuration, the integration of memory and logic functions can be achieved on a single chip, resulting in improved performance, power efficiency and cost savings. 3D parallel stack on memory over logic is favovarble to increase High Band Width Memory (HBM), but does not allow to completely break the memory wall. On the other hand, the 3D sequential integration flow would allow the fabrication of Logic immersed in memory systems that could improve near memory computing. This is because 3DSI authorizes an unique fine grain memory and logic partitioning that thereby reduces interconnection length between logic and memory. Additionally, the use of MRAM technology provides non-volatility and fast switching, which are highly desirable features in memory applications. [Refs]

3D integration is also promising is a method for integrating analog, radio frequency, and input/output functions on a single chip using a three-dimensional configuration. This approach can improve performance, power efficiency, and cost-effectiveness by allowing for the integration of functions on multiple layers. It also allows for the optimization of each function's characteristics by using different technology nodes. As example, the RF function, which includes the transmitter and receiver for wireless application, could be implemented using a technology node optimized for high frequency performance and power efficiency. The Analog function, which includes the power management circuit, could be implemented using a technology node optimized for improved linearity and low noise performance. By

using different technology nodes for each function, the overall performance and power efficiency of the device can be improved. [Refs]

True-3D VLSI (Very Large Scale Integration) refers to the integration of digital logic, memory, and interconnects in a three-dimensional (3D) configuration on a single chip. This approach allows for the stacking of multiple layers of transistors, memory, and interconnects, resulting in increased device density and improved performance. True-3D VLSI is particularly useful for digital applications such as high-performance computing. Clustered functional stacks is a way of organizing the different functions in a 3D structure for better performance, power efficiency and cost savings. This is possible thanks to an independent optimization of the different technologies and materials used at each level. [Refs]

All of these approaches have been actively researched and have been demonstrated to have great potential in achieving high integration density and high performance. However, there are still many challenges to overcome before 3D sequential integration is widely adopted in commercial applications. Amongst them, we can cite high cost, high thermal density, and the need for new packaging technologies. Overall, the state of the art in 3D sequential integration is rapidly advancing, with new research and development efforts focused on addressing the challenges and enabling the widespread adoption of this technology.

1.1.2 Monolithic 3D integration

3D integration refers to two types of 3D stacking: *the parallel integration*, in which the different layers are manufactured separately before being stacked together, and *the sequential or monolithic integration*, in which the layers are manufactured successively on top of each other on the same substrate as illustrated in **Erreur ! Source du renvoi introuvable.**. The main difference between the two options arises from the alignment process. For parallel integration, the alignment of the different layers occurs during the bonding process between the two plates. For the sequential integration, the alignment is, this time, achieved during the manufacturing of the second layer. This results in a higher precision for the latter approach, which is closer to a traditional back-end lithography step. As a result, the minimum spacing between contacts, which is dependent on the misalignment between the layers, is much smaller for 3D sequential integration. This allows for a higher density of contacts, with 10^8

contacts per mm^2 possible under 14nm design rules, resulting in several orders of magnitude higher density of 3D contacts that is limited to 10^6 contacts per mm^2 (fig.**Erreur ! Source du renvoi introuvable.**).

The increase in 3D interconnection for a sequential integration scheme is attractive for more than Moore applications, since it increases device density beyond device scalability itself. The implementation of different systems on separated layers is attainable thanks to the 3DSI high contact density. Its main interest is the elimination of costly trade-offs coming from the optimization of different devices on the same substrate by separating different functions on separate layers [Batude IEDM 17].

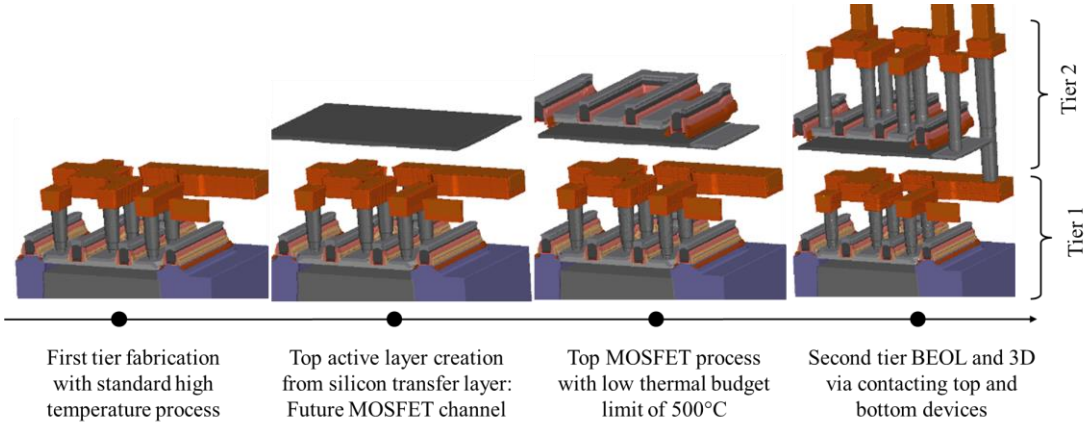


Figure 1-2 3D sequential integration process steps. After the fabrication of the first tier using standard high temperature process, the active layer is transferred above the finished back end. This crystalline silicon layer is used for the fabrication of the second tier using a low temperature process. Finally, both tiers are connected using 3D interconnections.

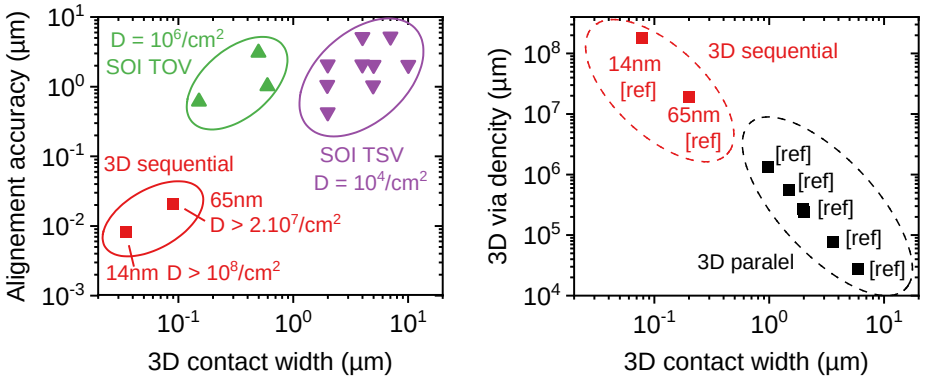


Figure 1-3 Alignment accuracy (left) and 3D contact density vs 3D contact pitch for different technologies obtained from parallel packaging and sequential integration. [Brunet16]

The first challenge in integrating the second tier using a sequential integration scheme is the formation of the silicon film used for the transistor channel. The most

consistent and performant solution is to transfer an SOI wafer onto the initial substrate through molecular bonding between two oxides. This method is a highly attractive approach as it can be performed at low temperatures (400ºC) while yielding defect-free material with exceptional thickness uniformity control [A. Mallik IEDM]. The substrate and buried oxide from the donor wafer are trimmed to the required dimensions to form the active region of the second tier. Although this process consumes a whole SOI wafer, which increases the cost of the second layer, it is currently necessary to obtain a high-quality monocrystalline region with minimal crystallography defects and low density of interface traps.

Another method is being developed that uses a bulk-Si donor substrate with embedded etch-stop layers for uniformity and thickness engineering, it does not consume the whole wafer and requires less mechanical and chemical trimming of the active layer. However, the drawback of this approach is that it results in a lower interface quality from the bottom channel that degrades MOSFET mobility [VLSI Brunet22].

Another suggested method is the formation of a polycrystalline active layer using silicon amorphous deposition that is crystallized using a nanosecond laser anneal [ref]. This approach is less expensive, but the polycrystalline channel produces devices with inferior quality and bigger variability. Alternatively, the channel could be formed from a crystalline seed on the second tier that is epitaxially grown through channels opened on the backend of the first tier [ref]. This approach preserves the crystalline structure from the seeds, but it occupies an area that cannot be used for device fabrication on both tiers, reducing the density of integrated components. Additionally, the recrystallization rate obtained with low temperature anneals may not be fast enough for recrystallization of large surfaces.

The second challenge in 3D integration is the fabrication of the second tier devices without damaging the lower layers. This requires careful management of the thermal budget during the manufacturing process, as high temperatures can damage the transistors and intermediate BEOL (back-end of line) metal lines below. Therefore, it is essential to establish the maximum thermal budget for the upper layer, taking into account the thermal tolerance of the lower layers, to prevent any degradation of their electrical performance.

Previous studies [Claire, IEDM, 2014 and Camila VLSI 2020] have investigated the effects of different furnace annealing temperatures on state-of-the-art FDSOI bottom CMOSFET transistors. The results showed that an annealing temperature of 500°C for 5 hours had no impact on the I_{ON} at a fixed I_{OFF}, DIBL, and for both NMOS and PMOS devices. However, metal gate work-function shift and silicide instability were observed at temperatures above 500°C. There were also no effects observed on the BEOL stability. The Time to Breakdown (TBD) of bottom tier transistors remained unchanged after a maximum annealing of 525°C for 2 hours, but began to degrade after a 2 minute anneal at 600°C. Yield analyses on 5 Mbit and 1 million flip-flops circuits also showed no yield loss for anneals up to 500°C. Therefore, it is possible to fabricate top-tier CMOS transistors above BEOL with 28nm by limiting the thermal budget to 500°C without requiring the integration of alternative materials for the Cu interconnections such as Co or W.

1.1.3 FD-SOI Transistor

Fully-Depleted Silicon-On-Insulator (FD-SOI) transistors are a type of MOSFET (Metal-Oxide-Semiconductor Field-Effect Transistor) device that are fabricated with a silicon film thickness under 20nm. The silicon channel layer is separated from the substrate by a buried oxide (BOX) layer. Due to the thin silicon layer, the channel becomes fully depleted, which allows a better electrostatic control of the channel and reduced fringe capacitances, leakage current and process variation from random dopant fluctuations. Additionally, FDSOI transistors have the ability for dynamic back biasing for threshold voltage modulation. As a result, FDSOI, a planar technology, is able to achieve the advantages of reduced silicon geometries and simplify the manufacturing process at the same time. This technology has been recognized as a leading silicon technology for applications involving low power, microwave and radio frequency.

These SOI substrates can be obtained using the SmartCut^{TM} technology developed by Soitec, which is based on the implantation of hydrogen ions through an oxide. The oxidized plate is then bonded above another plate and annealed, allowing the substrate to be separated at the level of the implanted hydrogen layer, resulting in a thin film of monocrystalline silicon on a BOX of variable thickness [Bruel95]. µ

1.2 LTB MOSFET FABRICATION

1.2.1 Description of the 28nm Integration process

The first step in the 3D sequential fabrication process begins with transferring an SOI wafer to the back end of the first tier (as shown in figure 1-4-a). The active region is then patterned on the silicon film using a step lithography process, followed by plasma etching. This patterning defines the geometry of the active region for each device, including the source and drain, and isolates different transistors to prevent latch-up.

The gate stack is then deposited over the entire wafer (**Erreur ! Source du renvoi introuvable.**-b). The gate oxide is the first layer of the stack and is composed of a thin interfacial silicon oxide layer (~1nm) followed by a thicker (~2nm) high permittivity (High-k) HfO_2 layer. The oxide layer improves the interface quality with the crystalline silicon film, while the thicker high permittivity HfO_2 layer reduces leakage current from direct tunneling between the channel and the gate while maintaining good electrostatic control. The stack of both materials can be interpreted as a single insulator with the same permittivity as silicon oxide (k=3.9) and an equivalent oxide thickness (EOT). The digital devices measured for this thesis have an EOT between 0.9 and 1.3nm.

The gate stack is followed by the deposition of titanium Nitrate (TiN) a metallic material useful for fixing the flat band voltage of the device and avoiding depletion of the polysilicon layer on strong inversion. Then, the formation of the polysilicon layer is performed with the deposition of an amorphous silicon layer in-situ, meaning the silicon material is deposited at the same time as the impurities for a homogenous doping profile. In other to crystalize the amorphous silicon layer and activate the in-situ dopants, a high temperature anneal is required. However, due to the thermal limitations from the sequential integration scheme, a classic high temperature anneal is not possible. In order to solve this problem, nanosecond UV laser annealing (UV-NLA) has been developed, which allows the wafer to be heated only in the upper layers (**Erreur ! Source du renvoi introuvable.**-c).

The short annealing time from the fast UV-NLA pulses strongly limits the diffusion of thermal energy from the annealed surface to the lower layers, potentially allowing temperatures above 1000°C to be reached for the upper layer without

exceeding the thermal budget limit for the lower layers. Previous studies showed that a temperature of 1200°C can be reached at the top level while keeping the temperature of the bottom level below 500°C. The non-vertical uniformity of this anneal ensures the integrity of the bottom layers while the upper layers can be annealed at high temperatures.

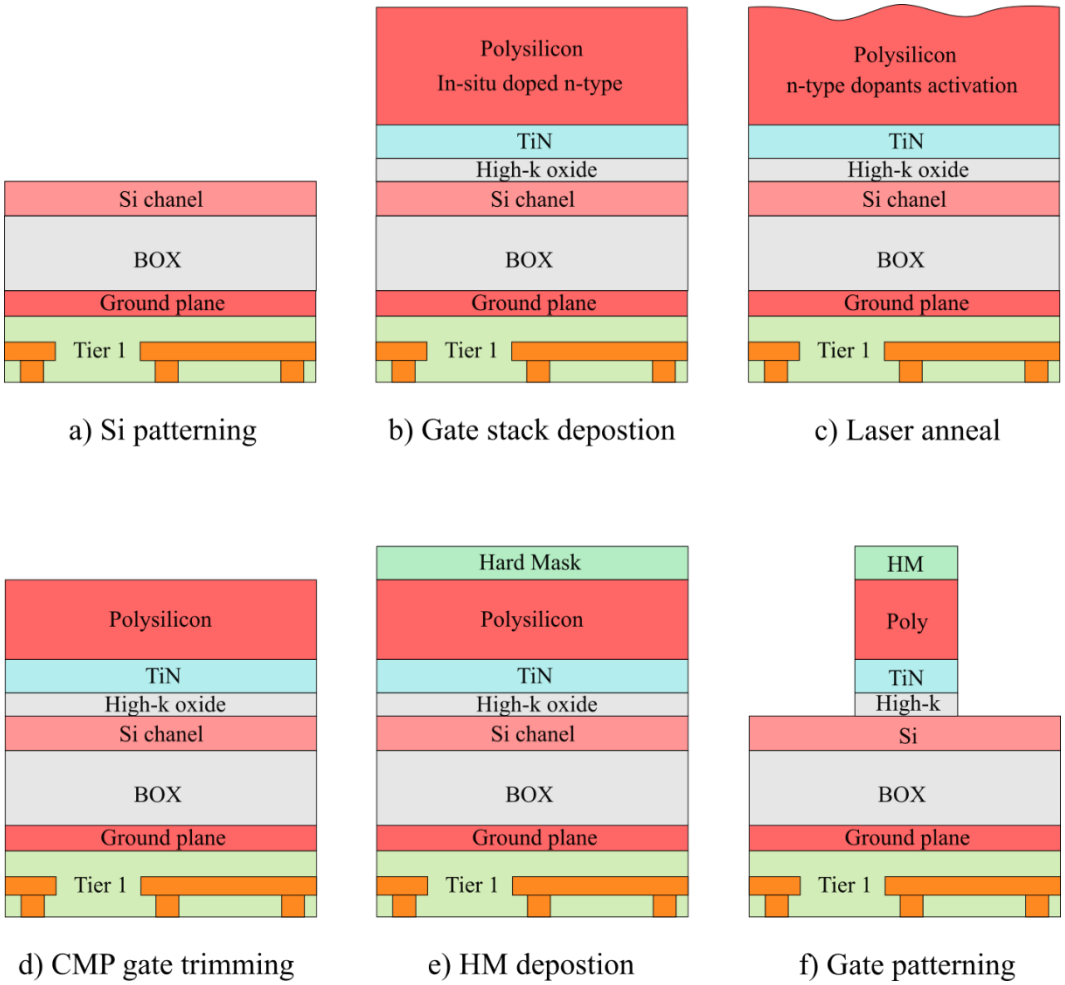


Figure 1-4 Top tiers low thermal budget fabrication steps (a to f). (a) Patterning of the silicon film to define the geometry of the active region of each devices. (b) Gate stack formation including the deposition of High-k and IL gate oxide materials, titanium nitrate metal gate and polysilicon layer. (c) Activation of polysilicon dopants using nano-second UV laser anneal. (d) Chemical-mechanical polishing to trim the annealed polysilicon surface. (e) Deposition of the hard-mask on top of the stack. (f) Etching of the stack to define the pattern of gate of each device.

It is important to note that the UV-NLA step also has drawbacks. In peculiar, the uniformity of the annealing depends on the geometry of the patterns present at the surface, because of the short wavelength of the laser. While the polysilicon is deposited above the whole wafer, the polysilicon thermal energy absorption is homogeneous and independent on the position. However, in the following steps, after the patterning of the devices, it is not yet a reliable method as it highly dependent on the transistor

geometry and heat absorption rate from different material stacks. Recent studies are been performed to improve the performance of UV-NLA annealing on patterned devices, but are still under development [ref].

Thus, nanosecond laser annealing is a promising technique for activating the gate dopants while limiting damage to lower components. On the other hand, the energy and number of pulses used on this process must be carefully optimized to control the activation temperature and avoid melting the polysilicon layer. The rugosity of the polysilicon layer is greatly increased after the LNA step, resulting in a non-uniform surface on the vertical axis. Therefore, the laser anneal is followed by a chemical mechanical polishing (CMP) that flattens the polysilicon surface and trims it to the required gate height (**Erreur ! Source du renvoi introuvable.**-d).

A hard mask (HM) composed of Silicon Nitrate (SiN) is then deposited above the trimmed surface (**Erreur ! Source du renvoi introuvable.**-e). This material is necessary to isolate the gate from the step of epitaxy of source and drain that will be performed latter and to improve the gate etching selectivity. Because the resist mask can degrade during the plasma etch, the resolution of the image patterned into the dielectric layer is reduced. This type of imperfect image transfer compromises the semiconductor device's performance. As high etch selectivity can be achieved between the hard mask layer and the over-coated patterned resist, image transfer imperfections can be avoided. The gate stack is finally etched to the gate geometries from the mask (**Erreur ! Source du renvoi introuvable.**-f).

After deposition and etching of the gate stack, the lateral part of the gate must be protected from the following epitaxy process. For this, a dielectric is first deposited and etched in order to form spacers that electrically isolate the gate from future source and drain accesses (**Erreur ! Source du renvoi introuvable.**-g). This spacer is created with the deposition of an insulator above the whole wafer surface followed by an anisotropic plasma etching that consumes the insulator material mostly in one direction. Next, the insulator is vertically etched until the hard mask and silicon film are cleared. Only the material deposited at the lateral of the gate stack remains at the end of this process step. On high temperature devices, the spacers is formed using the same material from the hard-mask, SiN, which has a permittivity equal to 7. However, the electrostatic performance of the transistor can be improved with a material with lower permittivity like SiCO. The main advantages of SiCO material are (1) the high

etching selectivity when compared to amorphous and crystalline silicon (2) a lower permittivity equal to 4.5 (3) its low temperature deposition (400°C), that makes it compatible with a LTB process deposition process [Benoit IEDM 15].

Because the silicon film is very thin (5nm - 10nm), a step of raising the sources and drains is then necessary in order to create the source and drain contacts. This step is accomplished through epitaxial growth. (**Erreur ! Source du renvoi introuvable.**-h). Because of the previous hard mask and spacer deposition, parasitic growth of Si above the gate stack that would lead to short the S/D and the Gate is not possible. On traditional high temperature epitaxy process, the crystalline film is growth through a selective silicon deposition at 750°C.

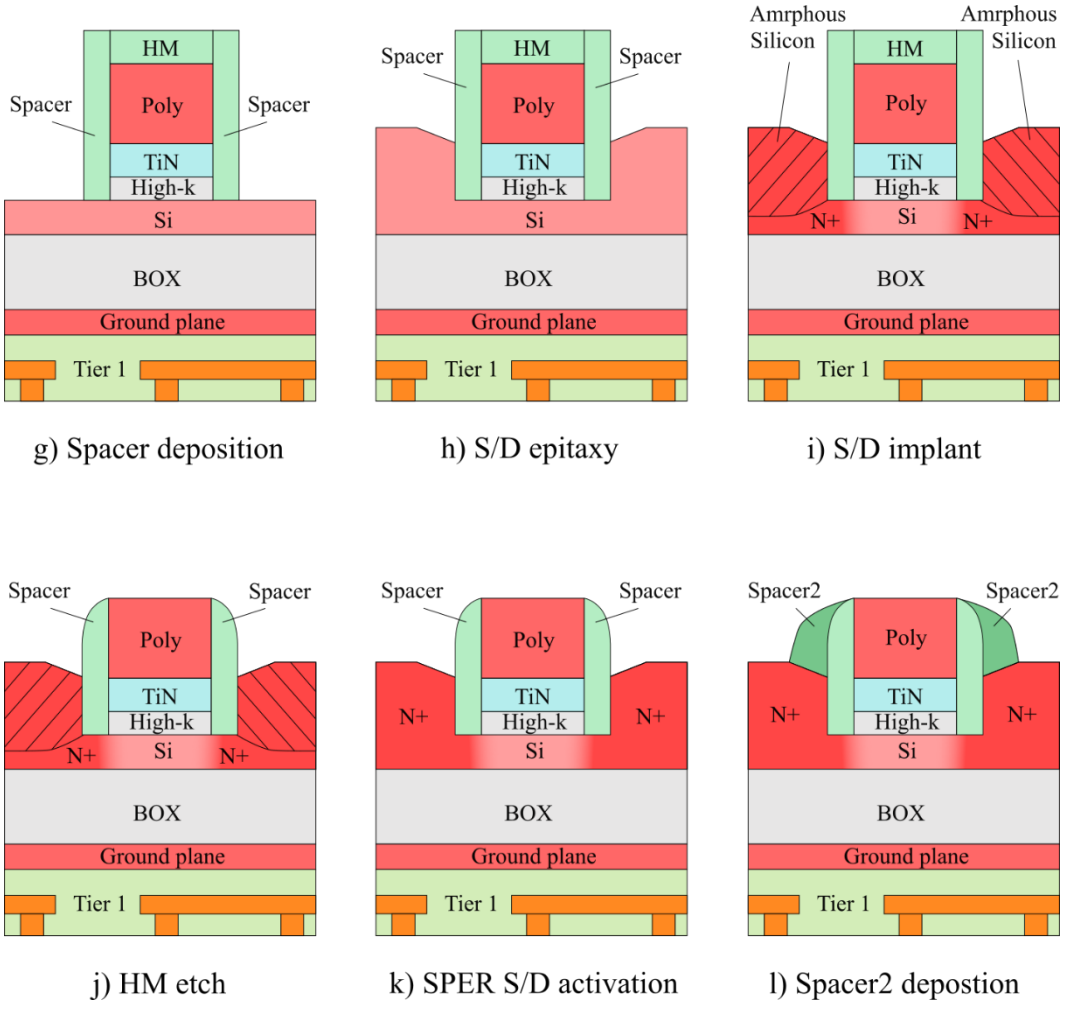


Figure 1-5 Top tiers low thermal budget fabrication steps (g to l). (g) Deposition of SiCO spacers to electrically insulate the gate and avoid parasitic growth during the epitaxies. (h) Source and drain epitaxies in order to improve access contact. (i) Source and drain doping and amorphization implant (j) removal of hard mask. (k) Activation of dopants from source and drain with SPER anneal. (l) Deposition of spacer2 to limit the reach of the silicidation step.

A new epitaxial process at was developed at LETI [Lu] where the temperature of two critical process were reduced to 500°C: the surface preparation and the epitaxial growth. The surface preparation is performed by a wet HF clean followed by an in-situ Siconi. The siconi process remove the surface oxide and is done at the same chamber from the epitaxy step. A Cyclic and Deposition Etch (CDE) was also developed for the low temperature process flow. The LTB epitaxy deposition phase is non-selective, where amorphous silicon is deposited above the oxide layers in addition to the crystalline region growth. Therefore, a deposition followed by an amorphous silicon etch is performed in sequence until the total epitaxy thickness is obtained.

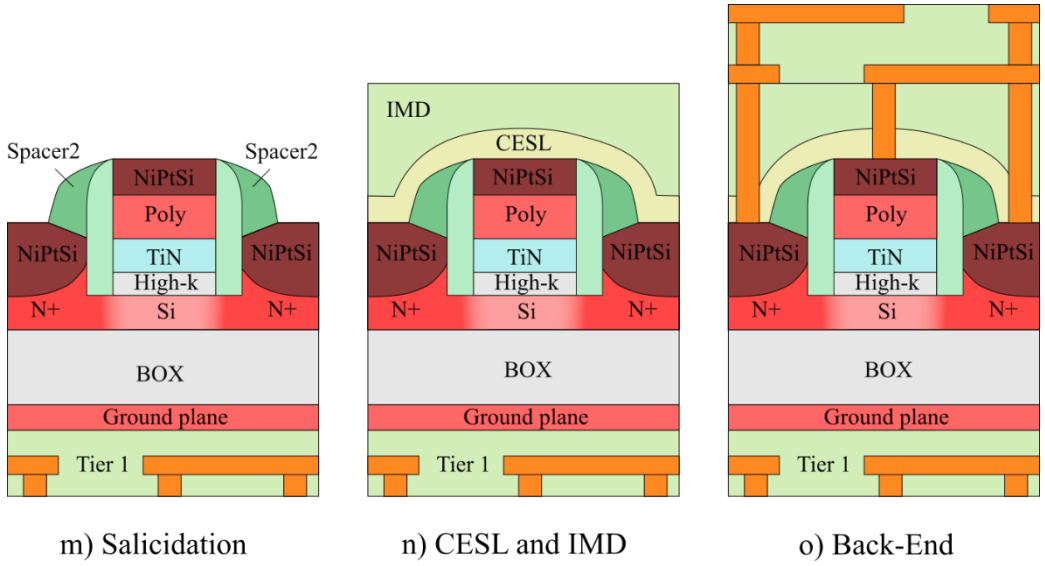
For high temperature process, the following step would be the source and drain dopants implantation and activation. At this stage, the choice of the dopant species will determine the type of transistor: phosphorus or arsenic implantation for nMOS or boron implantation for pMOS. Annealing is then carried out in order to provide the energy necessary for the dopants to substitute an atom of the semiconductor, activate them electrically, and recrystallize the region that becomes amorphous from the implantation. Once again, this step is not possible for the fabrication of top tier devices using the classic process because of the high thermal budget required for the activation of the dopants i.e. Spike anneal requires temperatures above 1000°C.

The alternative to high temperature annealing is the SPER technique, or Solid Phase Epitaxy Regrowth. The technique consists of recrystallizing an amorphized silicon layer from a monocrystalline layer (called the seed), allowing the crystalline orientation to remain unchanged. The amorphous layer is created either from the dopants or from germanium implantation (**Erreur ! Source du renvoi introuvable.**-i). For nMOS, the amorphization and doping is performed at the same step with phosphorous implantation. On the other hand, for pMOS, a pre-amorphization is performed by Germanium implantation, prior to the Boron implantation for doping. This is necessary, as the boron atom is not heavy enough to create a deep amorphization at the source and drain epitaxies.

Before the recrystallization of source and drain, the remaining SiN hard mask above the gate must first be removed. In order to achieve this, temporary SiO_2 spacers are deposited and etched anisotopically. This is required to protect the SiCO spacers during the removal of the hard mask. (**Erreur ! Source du renvoi introuvable.**-j). The Hard-Mask is then selectively removed typically in an H3PO4 solution, which has

a small effect on the SiO_2. Following the removal of the hard mask this temporary SiO_2 spacers are also etched.

After the hard mask removal, the SPER activation step is performed (**Erreur ! Source du renvoi introuvable.**-k). During the recrystallization, the atoms reorganize and the dopants integrate the silicon matrix in a substitutional position, becoming electrically active donor or acceptor dopants. A major difference between a low and high temperature activation lies on the dopant diffusion. During a high thermal budget process, the high temperature activation anneal allows the dopants to diffuse from the implantation region throughout the area underneath the spacers. However, no dopants diffusion happens with the low temperature SPER anneal. Therefore, the dopants must be placed directly at its final position under the spacer with a tilted implant.



m) Salicidation n) CESL and IMD o) Back-End

Figure 1-6 Top tiers low thermal budget fabrication steps (m to o). (m) Silicidation of contact regions of source, drain and gate.(n) Deposition of Contact Etch Stop Layer and Inter-Metal Dielectric layers above the finished device. (o) Fabrication of the back end layers

Following the activation of source and drain, a second set of spacers (Spacer2) is deposited and etched around the first spacers (**Erreur ! Source du renvoi introuvable.**-l). This step is necessary to distance the silicidation region from the channel in other to improve the access resistance. The silicidation is a formation of an alloy between the silicon and a metal and it is performed on the exposed silicon region of the source, drain and gate. The first step is the deposition of a NiPt alloy on the entire plate. Then a first anneal is performed so the silicon reacts with the metal to form the silicide NiPtSi. The anneal is followed by a selective wet removal of the unreacted metal and then by a second anneal to optimize the resistivity of the silicide

(**Erreur ! Source du renvoi introuvable.**-m). The silicide is an important step to reduce the source and drain access resistance of the transistor. In addition to the reduction of the local resistivity, it suppresses the schottky diode, initially formed between the metal and the semiconductor. This done by a reduction of the barrier height between the metal work function and the Fermi level of Si in the access region. [ref]

A Contact Etch Stop Layer (CESL) is then deposited to encapsulate the transistor and is followed by the deposition of an Inter-Metal Dielectric (IMD) (**Erreur ! Source du renvoi introuvable.**-n). The CESL is an insulator layer that improves the control of the etch depth for the contacts by reducing its dependence on the on the local topography. In essence, the CESL material has a low sensibility to the solution used for IMD etch. A first contact etch will remove the IMD material until the CESL layer. A second careful etch step is performed on the CESL with a different solution to reach the silicide region. Etched cavities are then filled by a Ti/TiN barrier and tungsten material (**Erreur ! Source du renvoi introuvable.**-o). In addition to the improved etching control, the CESL can introduce mechanical compressive or tensile stress on the transistor in order to improve the carrier mobility in the channel.

1.2.2 Extension-first vs Extension-Last

As mentioned before, the issue with the SPER anneal is the non-diffusion of dopants at low temperatures and its dependence on the amorphization region for the activation of dopants. Yet, it is not trivial with a single implant to place the dopants under the spacers at their right position. The high implantation energies generates stronger variability and the spacer thickness is limited by the maximum distance traveled by the dopants. In addition, due to the activation dependence on the recrystallization, only the dopants present in the initially amorphized layer are activated during SPER.

Co-optimizing the implantation step is a challenging task due to the conflicting effects that arise. If the implantation is too strong, it will result in the absence of a crystalline silicon seed, which is necessary for SPER recrystallization. On the other hand, if the implantation is too low, it will lead to a limited amount of dopants under the spacers, increasing the source and drain access resistance. Despite these challenges, there is limited understanding of the actual need to amorphize the region under the

access in order to achieve a high activation level. Recent studies suggest that even a low-temperature anneal at 500°C may be sufficient to activate the dopants under the spacers due to their proximity to the oxide interfaces of the spacers and the BOX.

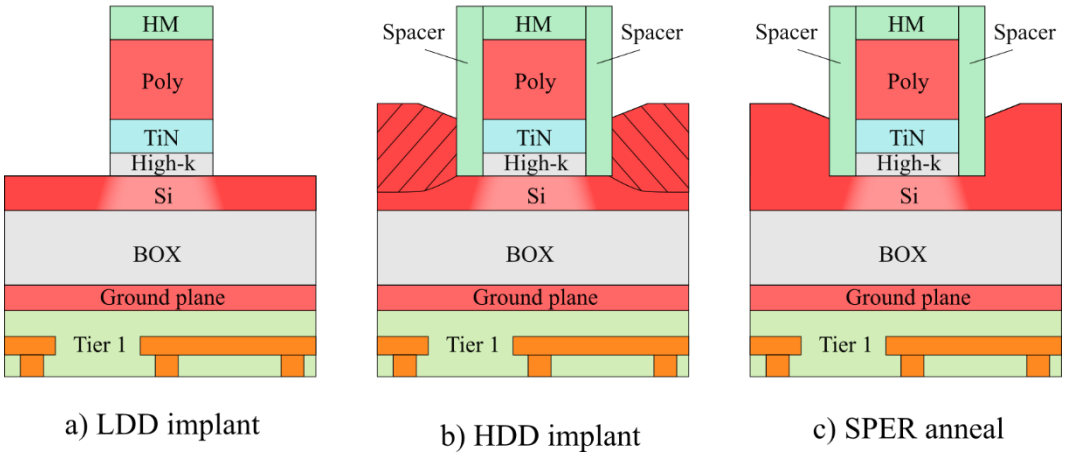


Figure 1-7 Top tiers low thermal budget fabrication steps on an extension first integration scheme. (a) LDD Implantation of source and drain before the deposition of gate spacers. (b) HDD Implantation after spacer deposition and epitaxy of source and drain. (c) Activation of dopants with SPER recrystallization anneal.

The process flow presented so far is called extension-last, where the source and drain implantations are performed after the source and drain epitaxies. Another potential solution exists to carefully optimize the junction based on a so-called extension-first integration flow. In that case, the implantation is performed, prior to the raising of the source and drain or even before the formation of the gate spacers. A first LDD (lightly doped) implantation before the spacer deposition or with a thinner spacer (2-3nm) allows placing dopants near the channel with smaller implantation energies (**Erreur ! Source du renvoi introuvable.**-a) [ref Pasini16b]. Then, a second HDD (Highly doped) implantation is performed after thicker spacers' deposition and epitaxy of source and drain (**Erreur ! Source du renvoi introuvable.**-b). The HDD implants are then activated with the recrystallization from the SPER anneal (**Erreur ! Source du renvoi introuvable.**-c). This last integration has been demonstrated [Pasini16] on a 14FDSOI technology from STMicroelectronics with a maximum thermal budget of 600°C. The device presented electric performances close to a reference at high temperature on nMOS and pMOS.

1.2.3 Low vs High thermal budget process

Erreur ! Source du renvoi introuvable. summarizes the several integration steps required for the fabrication of low temperature silicon MOSFET transistors.

Firstly, the deposition of the gate oxide stack and the reliability anneal that are normally performed between 800°C and 900°C are reduced to 500°C using a method developed at LETI [ref?].

In addition, the deposition of the spacers that are classically fabricated using a 630°C silicon nitride deposition are switched to a low temperature 400°C SiCO deposition , which also benefits of a lower permittivity. The epitaxy of source and drain is another step that was re-developed specifically for this technology where the deposition of crystalline silicon and etching of amorphous layers are done separately in cyclic steps. Finally, the activation and diffusion anneal that is normally done at temperatures above 1000°C is replaced by the SPER amorphization and recrystallization anneal that allows for a high activation of impurities under 500°C in the epitaxial region of source and drain.

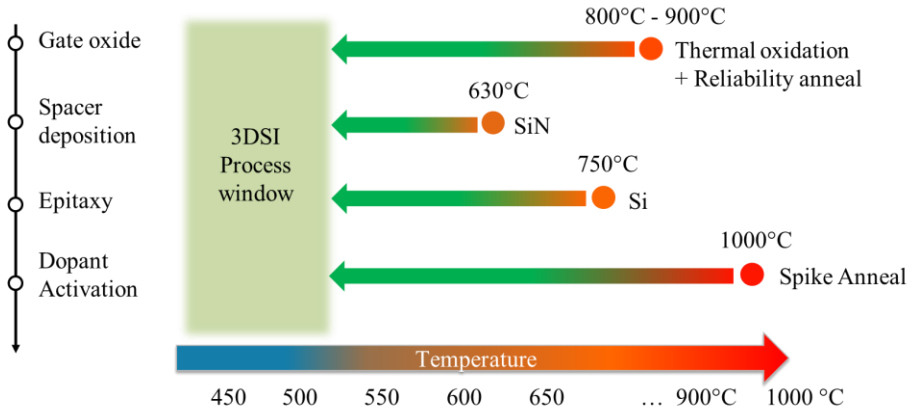


Figure 1-8 Representation of key technological steps that are re-developed at LETI for the low thermal budget integration. The deposition of the gate stack, formation of spacers, source and drain epitaxies and activation anneal are performed under 500°C to avoid degradation and guarantee the stability of transistors and interconnections from lower tiers.

The studies performed during this thesis focus on two of those re-developed processes. Firstly, the impact of the low temperature SPER technique on the S/D junction properties is fully characterized. In particular, the thermal activation of Bore and Phosphorus impurities implanted in the non-recrystallized region and its impact on the access resistance and short channel effects is explored. In addition, the quality of the SiCO oxide used for the formation of the spacers is also studied on low thermal budget MOSFET devices. This has been done by analyzing the trapping behavior of the oxide and its detrimental effects on device performance. Both studies resulted in the development of novel characterization methods that have been used to provide a

deeper understanding of the effects of those new fabrication methods on the electric behavior of the devices.

1.3 PURPOSE OF THIS THESIS

The objective of this thesis is to explore the effects of low temperature integration process developed for the fabrication of 3D-SI FD-SOI MOSFET transistors and to propose optimization paths for the performance improvements of digital technologies in the high frequency domain. The studies were performed using low temperature devices fabricated at LETI and are currently compared to devices fabricated using a classic high thermal budget. All the studies have been made on single tier MOSFET wafers fabricated within a low thermal budget, representative of the budget required for the processing of the top tiers of a 3DSI integration. The objective is therefore to characterize the effects of the low temperature process on the electric behavior of the device and improve its fabrication at low temperature.

This manuscript is divided in 3 parts following this introduction. The second chapter focuses on the optimization of the junction position on an extension first integration scheme. A novel method for the extraction of the junction profile on ultra-thin FD-SOI MOSFET transistors is developed and validated against classic high thermal budget devices. Then, the method is used to determine the level of activation obtained from different implantation conditions on NMOS and PMOS transistors, and to correlate it to their electric behavior. Based on the obtained results, new integration schemes are proposed to improve access resistance, reduce variability and parasitic capacitances of future devices.

The third chapter focuses on the SiCO spacers developed for the LTB technology. If the main material properties of the SiCO oxide, such as conformal deposition, etching robustness, were well known at the beginning of the thesis, only few studies were performed so far on its electrical properties. We therefore choose to determine the main electrical properties of this material, and, how they affect the device performance. For this, two new ultra short capacitance measurement methods are developed that allowed to fully characterize the complex trapping behavior of this material. The method were also applied on high temperature oxides used on the development of thermal budget devices. Finally, the effects of hydrogen and high temperature anneals were characterized using the same methods.

The fourth chapter describes the interest of 3DSI for RF application. The fey RF FoMs (f_t and f_{max}) are extracted on our LTB transistors and are compared to the HTB

counterpart. A full ac modeling of LTB device in the RF domain is also provided to determine the process leverages for a further boost of RF device performance. As conclusion, a path for improving the performance of the transistors for high frequency applications is proposed where the devices could be used for a specialized high frequency layer for the development of transceiver SoCs.

Chapter 2: Characterization and Optimization of Junctions of FD-SOI transistor

Traditionally, during the MOSfet integration process, dopants of the source and drain are introduced to the silicon substrates by ion implantation. In this method, impurities are ionized into a plasma and then are accelerated by an electrical field under a strong vacuum into the substrate surface. When an ion strikes the substrate, it collides with atoms in the lattice, taking a random scattered path before being completely stopped. After the implantation, a vertical Gaussian profile of impurities concentration is obtained that depends on the beam energy, dose and the implanted species. The implanted ions will not necessarily occupy substitutional positions in the silicon lattice. Meaning that it will not be electrically active on the semiconductor gap. In addition, the collision with the substrate atoms can create localized crystalline defects generating silicon vacancies and interstitials. Therefore, a post-implant anneal is necessary to activate the dopants by substituting the impurities atom into crystal lattice sites and to repair any damage done to the silicon crystal. This high temperature short time anneal is also responsible for the diffusion of the impurities inside the crystal. This mechanism is important for the fabrication of MOSfet devices because it allows dopants to be electrically active and to diffuse under the gate spacers required to form a low resistivity path connecting the source and drain access to the channel. [Fundamentals of semiconductor Manufacturing and process control].

For 3D sequential integration, a low thermal budget (<500°C) is required for the fabrication of devices of the top layers. This restriction prohibits the use of traditional high temperature methods for activation and diffusion of dopants. Therefore, on LTB devices the activation step is performed by a different method called Solid Phase Epitaxial Regrowth (SPER). It is based on a layer-by-layer recrystallization phenomenon that happens on amorphous/crystalline (A/C) silicon interfaces that can be triggered at lower temperatures. During the crystal regrowth, the dopants impurities present on the substrate are incorporated into the lattice where they become electrically active.

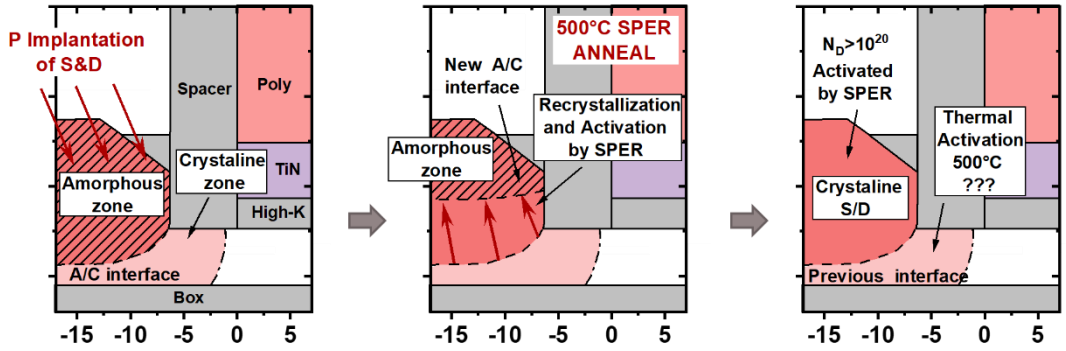


Figure 2-1 Sketch illustrating the principle of the SPER recrystallization anneal on n-type LTB devices.

The strategy for dopants activation on LTB devices is presented on **Erreur ! Source du renvoi introuvable.**. It consists in the creation of an A/C interface on the source and drain epitaxial regions and the subsequent activation of the impurities of the amorphous zone by crystal regrowth. The transformation from a crystalline to an amorphous phase is performed by ion implantation. If the implantation is strong enough, it can trigger collision cascade effects near the surface that transforms the region from crystalline to amorphous. On n-type devices, this is performed directly by the phosphorous implantation. Therefore, the formation of the amorphous region and the introduction of dopants is performed on the same step. On the other hand, for p-type devices, a pre-amorphization step must be performed. This is necessary because boron atoms are not able to break the crystal, because of their lighter atomic mass. Then, boron ions are introduced to the substrate with a second ion implantation step.

After the SPER, a high amount of dopants of the amorphous region can be activated, throughout the recrystallization process. Because the dopants are incorporated into the lattice in a metastable equilibrium, their initial concentration can largely exceed the solid solubility limit of these impurities on silicon at low temperatures. These solubilities have been determined experimentally by [Passini] and correspond to an activation level at 600°C of 10^{21} cm^{-3} for phosphorous and $9 \cdot 10^{20}$ cm^{-3} for boron and at 500°C of $6 \cdot 10^{20}$ cm^{-3} for phosphorous and $5 \cdot 10^{20}$ cm^{-3} for boron, respectively. On the order hand, the activation of dopants confined in the region that remained crystalline after the implantation are limited by the thermodynamic equilibrium solubility of the specie. The solubility of dopants at crystalline silicon can be modelled by an Arrhenius behaviour, and it is much smaller

at T=500°C ie approximately $10^{19}cm^{-3}$ and $10^{18}cm^{-3}$ for phosphorous and boron respectively [Passini].

The low theoretical solubility limit at low temperatures makes challenging the formation of the junction using the SPER technique. On one hand, a depth amorphization is necessary to properly activate the dopants under the spacer region during the SPER. On the order hand, it is important to preserve a crystalline seed of around 2nm on the bottom of the silicon layer to trigger the recrystallization process. In addition, the preamorphization followed by recrystallization step also induces End of Range (EoR) Defects on the silicon film. These defects are interstitial type dislocation loops that are formed under the crystalline/amorphous (A/C) interface. They originate from the presence of highly saturated interstitial Si atoms that become dislocated after the implantation. [C. Bonafos,]. The presence of these defects can negatively affect the electrical performances of the transistors. They reduce the activation of dopants near the (A/C) interface, and increase the gate induced drain leakage (GiDL) by trap assisted tunnelling [ref].

A second solution that have been explored is the activation of the access region using the Nano-second laser annealing (NLA). This method allows highly increasing the temperature on the top layer structures without affecting the bottom tiers. However, it imposes several challenges; Firstly, it can create a highly uneven silicon surface that must be treated using a chemical-mechanical planarization (CMP) step. In addition, the increase in temperature depends on the dimension and capability for heat dissipation of the targeted structures at the surface. Therefore, this anneal can induce different effects between small and large devices.

The last integration scheme is yet the most reliable solution for the LTB process. Recently, devices presenting high FoM and low access resistance were demonstrated [VLSI] without amorphization of the region under the spacers, and, despite the theoretical low solubility limit. Therefore, a more in-depth study on the maximum solubility and the activation phenomena of dopants on ultra-thin crystalline silicon films must be performed.

In this chapter, the junction position and the density of dopants that are activated with the LTB SPER method are studied. To evaluate the efficiency of activation under the spacer region, we developed a novel technique called CV-AJP to extract the

junction active profile in FDSOI devices using capacitance measurements. Then a study of the correlation between the electrical performances of the device and the junction position is performed for NMOS and PMOS devices. Finally, based on the results, a new source and drain implantation solution for the extension first integration flow is proposed.

2.1 ELECTRICAL CHARACTERISTICS OF THE ACCESS REGION

2.1.1 Impact of the junction profile on device electrical features

Erreur ! Source du renvoi introuvable. presents the effects of the junction profile on the Ion/Ioff FoM for a 27nm MOSfet transistor. The device was simulated for several Gaussian profiles with different values of overlap and density of dopants. A device with an overlapped junction will suffer from stronger Short Channel Effects (SCE). The threshold voltage in saturation (Vt_{sat}) decreases and the subthreshold slope (SS) increases. On Ion/Ioff FOMs, the change of the overlap position (points with same colour) produces an Ion/Ioff variation, similar to one produced a variation of the channel length. On the other hand, the doping density has a direct effect on the access resistance. Increasing the doping level increases the carrier density in the access region, that, in turn reduces the access resistance, and the slope of the Ion/Ioff characteristics.

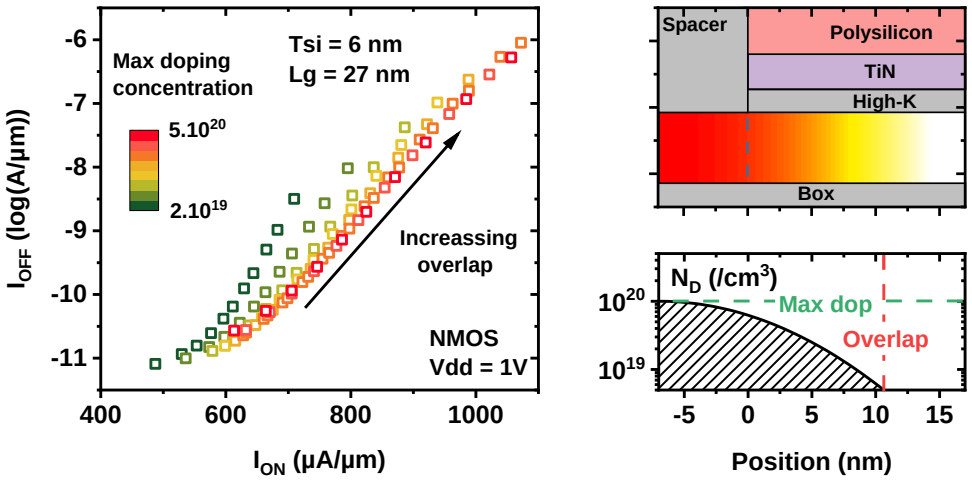


Figure 2-2 TCAD simulation of a 27nm channel length MOSfet at different Gaussian profiles for the source and drain junctions (Left) Ion/Ioff at $V_{dd} = 1V$ FoM for each profile condition (Right) The Gaussians profile parameters are the maximum amount of dopants and the overlap position. The overlap is arbitrarily set at the position where the doping density is equal $5 \cdot 10^{18}$ cm^{-3}.

Erreur ! Source du renvoi introuvable. also illustrates how SCE are dependent on both the density of dopants and on the overlap position. This can be better visualized at **Erreur ! Source du renvoi introuvable.** where the values of saturation Threshold voltage (Vt_{sat}), Drain induced Barrier lowering (DiBL) and Subthreshold slope (SS) are extracted for each profile condition. Increasing the overlap also increases the short channel effects regardless of the junction profile. Nevertheless, devices containing a

smaller concentration of dopants in the access region also exhibit a reduced DiBL and SS at same fixed overlap.

The same analysis can be performed from gate capacitance simulation. **Erreur ! Source du renvoi introuvable.** illustrates how the depletion capacitance (C_{dep}) at $V_G = 0V$ changes as the function of the doping profile. **Erreur ! Source du renvoi introuvable.** (Left) reports the simulated capacitance for distinct junction conditions. For the first case, an overlapped profile containing a small density of dopants (a) is compared against an overlapped profile with a high density of dopants (b). Even though the capacitance curve is very distinct on depletion, both profiles present the same value of C_{dep} at $V_G = 0V$. Alternatively, a second case shows two profiles (c and d) with the same overlap but with different doping densities. The capacitance in depletion is different between c and d despite the same junction position indicating that the depletion capacitance by itself is a misleading parameter to determine properly the overlap position. Finally, **Erreur ! Source du renvoi introuvable.** (Right) present the values of C_{dep} obtained from different access profiles. Similarly to the IV simulations, the capacitance is also dependent on both the overlap position and density of dopants near the junction.

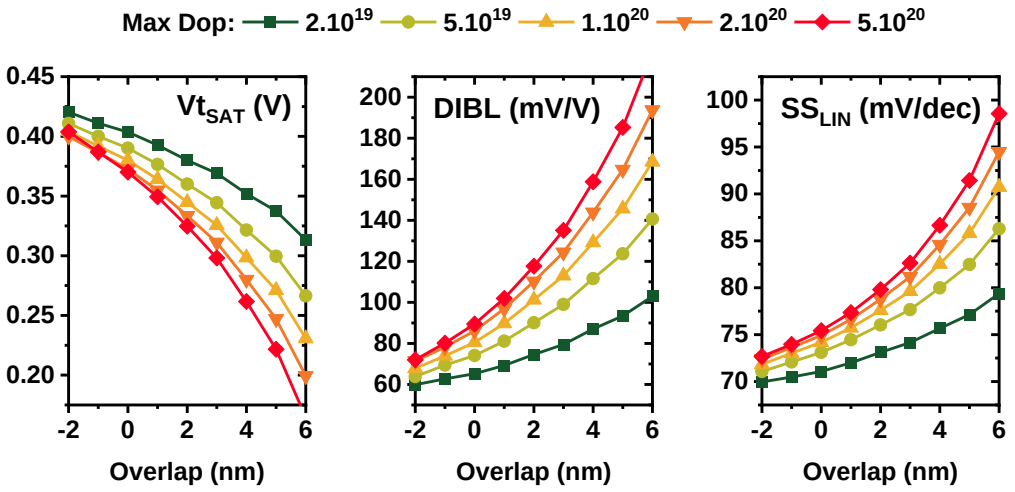


Figure 2-3 Standard device parameters extracted from IV TCAD simulation of a 27nm MOSfet for different S/D junction profiles. (Left) Vt_{sat}. (Middle) DiBL. (Right) Linear SS.

The fact that electrostatic parameters like SS and C_{dep} are not exclusively dependent on the junction position is explained by a difference between the metallurgic junction and the carrier junction. The metallurgic junction is related to the amount of dopants on the access. Traditionally, it is defined by the exact position where the

density of acceptors equals the one of donors. However, this definition becomes less useful in the most advanced technological nodes, when the channel is much less doped. Moreover, there is no wide accepted minimal value of dopants for the metallurgic junction on intrinsic channels so on this manuscript we will use the density value of 10^{19} cm^{-3} to define the metallurgic junction.

The true electrical junction, on the other hand, is not only related to the profile of dopants but also to the electrical characteristics of the device, such as bias conditions for instance. Basically, in MOSfet devices, a depletion region is formed on both sides of the metallurgic junction that depends on the junction parameter (dopant concentration and abruptness) as well as the control produced by the gate through V_G. This depletion in the access region will strongly depends on the values of V_{GD} and V_{GS}, and therefore on the corresponding regime for the transistor i.e. channel depleted $V_G < V_T$ or inverted $V_G > V_T$. The device electrostatic is directly linked to the variation of the depleted zone in the access region since it corresponds to a conductive limit of the access in depletion regime. In this manuscript, the electrical position of the junction will be defined as the point where the free carrier density equals to the doping density.

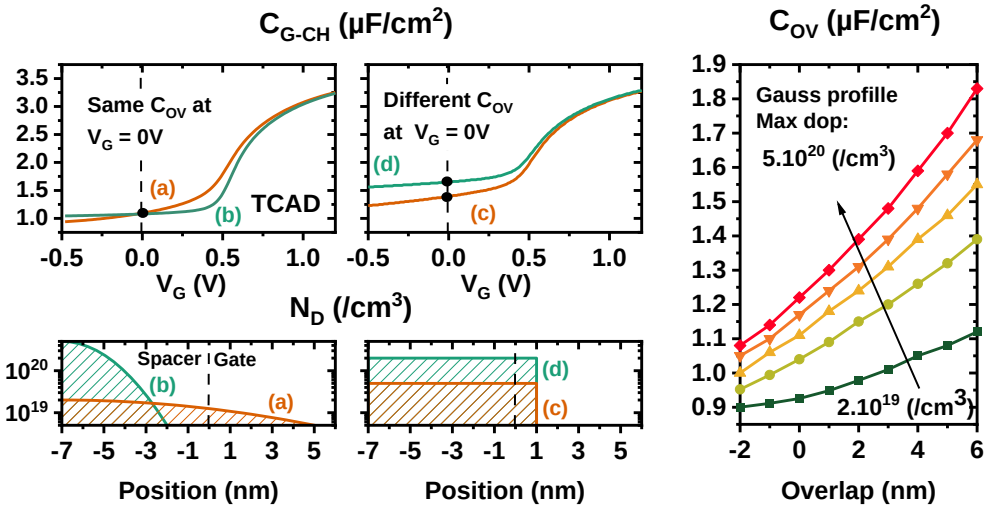


Figure 2-4 CV parameters extracted from TCAD simulation of a 27nm MOSfet at different profiles. (Left) CV curve obtained from distinct doping profiles presenting the same value of C_{dep} for profiles with different overlap positions and different values of C_{dep} for profiles containing the same overlap position (Right) C_{dep} for several profiles.

Erreur ! Source du renvoi introuvable. illustrates what is the real difference between metallurgic and electrical junction based on free carriers density, and its impacts on device capacitance. **Erreur ! Source du renvoi introuvable.** (left) presents the density of free carriers at two different gate biases chosen in the depletion

regime of the transistor. When V_G is reduced, a depleted region is formed on the access region beneath the spacers shifting the conductive region in the direction of the S/D. This phenomenon corresponds to a variation of the carrier junction position and a reduction of the inner and outer field capacitance. Obviously, the shift is going to be dependent on the density of dopants on the access. The lower is the doping, the larger is the volume that must be depleted in the access to counter the increase in the electrical field induced by the gate. This also induces a stronger variation of the capacitance value in depletion (**Erreur ! Source du renvoi introuvable.** c and d). Therefore, C_{dep} is not constant over V_G, its absolute value depends on the carrier junction position and its variation depends on the density of dopants.

For most of the cases, the effective channel length (L_{eff}) is the preferred metric to evaluate the overlap. Nevertheless, determining the effective channel length becomes challenging in very short channel devices. The several methods to extract L_{eff}, from the subthreshold slope, gate capacitance and the channel resistance over the device length become less appropriate for devices with a very small gate length. The reasons for that are manyfold. (1) the mobility changes from long to short devices (2) the access resistance is not negligible and (3) fringe parasitic capacitances are no longer constant as function of the physical length.

Beyond the extraction of the effective channel length from electrical parameters, there are physical-mechanical characterization methods that allows measuring the density of dopants directly from the device. Most notably, the scanning spreading resistance microscopy (SSRM) and the off-axis dark field electron holograph are the most promising techniques for the extraction of the profile.

The SSRM is an advanced atomic force microscope (AFM) characterization method that uses a small conductive tip to measure the local spreading resistivity. The density of dopants is then extracted by the direct correlation between carrier's density, mobility and resistivity. It has a spatial resolution typically of 1-3nm and can even perform 2D mapping of the dopants density [refs].

Off-axis electron holograph involves measuring the phase shift of an electron wave passing through the sample using a Transmission Electron Microscopy TEM setup. The phase shift of the incident wave is expressed as function of the inner electric potential of the specimen. Then, the junction position is extracted from the position of

a step in potential induced by the variation of dopant density. This technique presents very interesting results for sub µm bulk devices but fails to provide meaningful resolution on ultra-thin film advances FDSOI nodes. Additionally, SIMS and atom probe tomography are also very valuable but they measure only the total quantity of dopants atoms and not the electric active ones. [refs]

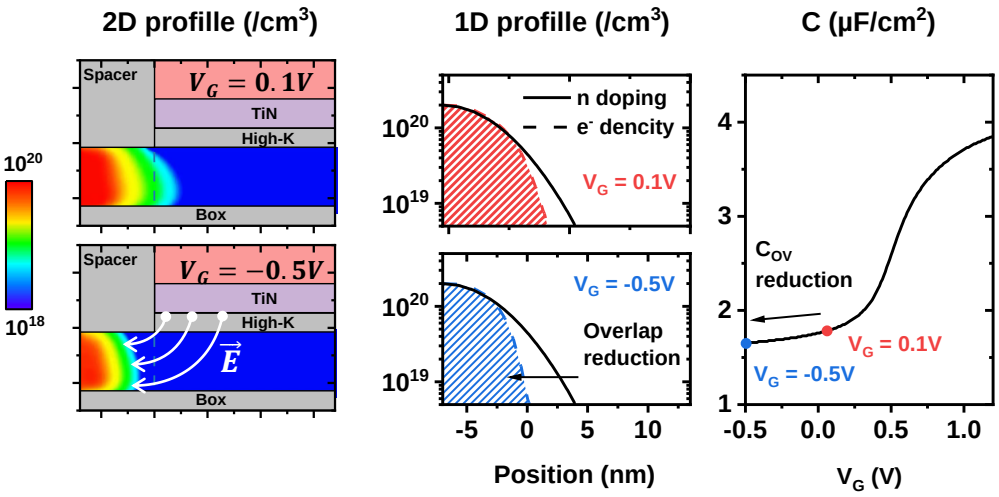


Figure 2-5 Illustration of the depletion of source and drain dopants on depletion regime ($V_G = 0.1V$ and $V_G = -0.5V$). (Left) 2D and (middle) 1D Carrier's density at the source for two gate biases on depletion with $V_D = V_S = 0V$.

Nevertheless, those sophisticated physical-mechanical methods are expensive. In addition to the use of an AFM or TEM setup, they requires a tedious preparation the sample consisting in polishing, cleaving or milling. Moreover, none of those methods provides the necessary accuracy to study a device with 27nm gate length and 7nm silicon film thickness. Therefore, in the next session, we present a novel methodology for an accurate extraction of the source and drain doping metallurgic profile on UT-FDSOI directly from the electric measurements.

2.2 ACTIVE JUNCTION PROFILE EXTRACTION METHODOLOGY

In this part, a novel methodology, called CV-AJP for Active Junction Profile is proposed to extract the source and drain junction profile based on CV measurements and modelling. In essence, from the model it is possible to obtain the overlap position that gives the value of measured capacitance at for each V_G bias in depletion. From that, we obtain the depletion junction position over gate bias. The density of dopants is then obtained from the variation of this parameter over V_G. For that, a precise extraction and modeling of the parasitic capacitances of a UT-FDSOI MOSFET device is prerequisite.

2.2.1 Extraction of parasitic capacitances

The first step of this method is to separate the parasitic components form the measurements using the variation of inversion capacitance over device length. **Erreur ! Source du renvoi introuvable.** presents the different capacitive components of the device on different regimes. The inner fringe capacitance (C_{if}^{w}) is related to the electrical field traversing the gate oxide into the depleted silicon film until the junction and it is screened out by the formation of the channel in strong inversion. The outer fringe capacitance (C_{of}^{W}) corresponds to the zone of the spacers between the lateral part the gate and the access. It is important to note that the value outer fringe capacitance in underlapped devices is different from inversion to depletion regime. Because the channel also inverts under the spacers, the value of C_{of}^{W} increases from depletion to inversion regime. As illustrated at **Erreur ! Source du renvoi introuvable.**, when the channel is completely inverted, the value of C_{of}^{W} is practically independent on the overlap. The oxide capacitance C_{ox}^{S} exists only on inversion regime and corresponds to the coupling between the gate and the channel. The box capacitance C_{box}^{W} is related to the coupling between the silicon channel and the substrate through the box. The (w) and (s) superscript indicates that the capacitance is given by width or surface length respectively; i.e. the total value is divided by the device width or surface. In addition, two capacitive components are independent of gate and drain bias V_G & V_D. C_{2D}^{w} denotes the 2D parasitic capacitance due to coupling between the gate and the source and drain and proportional to the device width. Note that C_{2D}^{w} is itself the sum of two contributions: (1) C_{G-Ctc}^{w} due to coupling between the gate and the contacts through the thick gate dielectric (2) $C_{G-S/D}^{w}$ the capacitance between gate and epitaxial region.

C_{3D} denotes the 3D fringe capacitance due to coupling between the gate and the source and drain contacts and back-end interconnections.

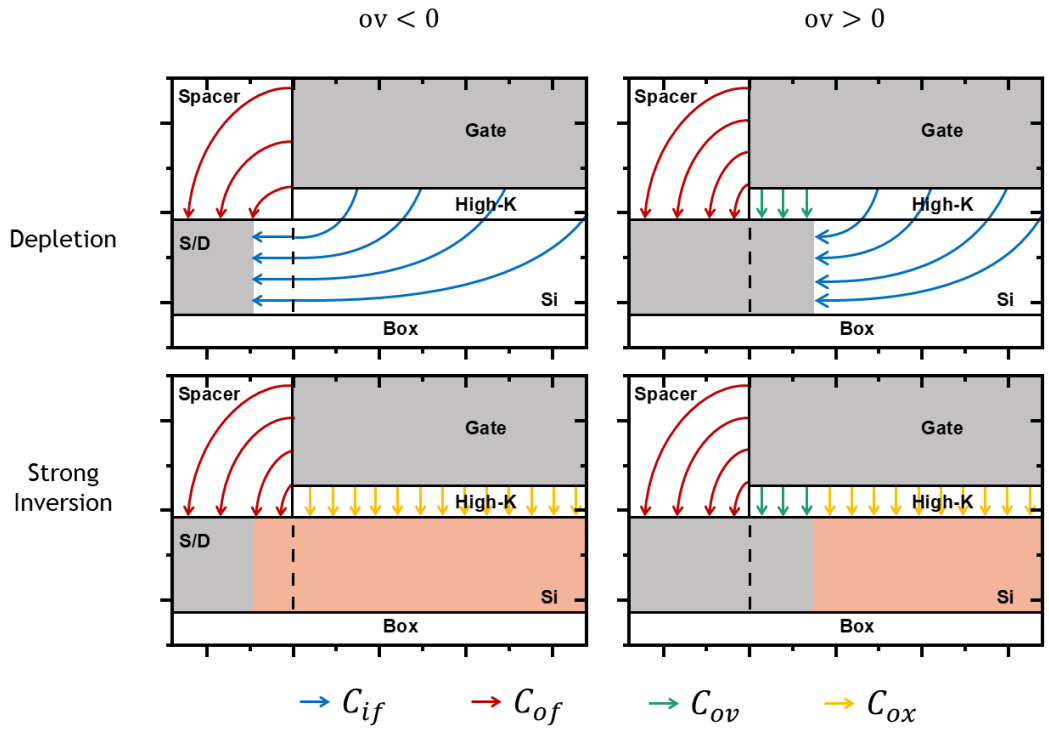


Figure 2-6 Sketch illustrating the different capacitive components that are dependent on the device bias (C_{if}, C_{of} and C_{ov}). The capacitive components are illustrated for overlapped (top) and underlapped (bottom) devices at strong inversion (right) and on depletion (left) regimes.

The total capacitance at the extreme of both regimes (C_{dep} for $V_G \ll V_T$ and C_{inv} for $V_G \gg V_T$) can be represented by the sum of the components presented on equations 2-1 and 2-2. Those equations consider that the devices capacitance is measured only between gate and access. As the box is connected to the ground and not to the capacimeter, C_{box} component does not contribute to the CV measure and is therefore not considered on equation 2-2

$$C_{inv} = WL_G C_{ox}^S + 2W\, C_{of}^W(ov = 0) + WC_{2D}^W + C_{3D} \quad (2\text{-}1)$$

$$C_{dep} = 2WC_{if}^W(L_G) + 2W\, C_{of}^W(ov) + WC_{2D}^W + C_{3D} \quad (2\text{-}2)$$

Then the methodology then consists in:

1. Plotting the measured capacitance in depletion C_{dep} vs W to extract the external gate contact to S/D contact capacitance C_{3D} (see Fig 2-7 (a))

2. Extracting C_{ox} and $2C_{of}(ov = 0) + WC_{2D}^W$ by plotting the inversion capacitance $C_{inv} - C_{3D}$ versus gate length L_G . C_{ox} is directly proportional to

the slope whereas $2C_{of}(ov = 0) + C_{par}$ is the interception at $L_G = 0$. (see Fig. 2-7) The value of C_{inv} is obtained at a fixed overdrive bias ($V_G - V_T$).

3. Deducing the access fringe depletion capacitance C_{ga} without *bias independent* parasitic components (eq. 2-3).

$$C_{ga}^{W} = \frac{C_{dep}(L_G) - C_{inv}(L_G = 0)}{2W} = C_{if}^{W}(L_G) + C_{of}^{W}(ov) - C_{of}^{W}(ov = 0)$$
$$= C_{if}^{W}(L_G) + \Delta C_{of}^{W}(ov) \qquad (2\text{-}3)$$

The $\Delta C_{of}^{W}(ov)$ component arises from the variation of the outer fringe components between inversion and depletion regimes. In addition, because C_{of}^{W} is maximized at strong inversion, the value of ΔC_{of}^{W} will always be zero for overlapped and negative for underlapped devices.

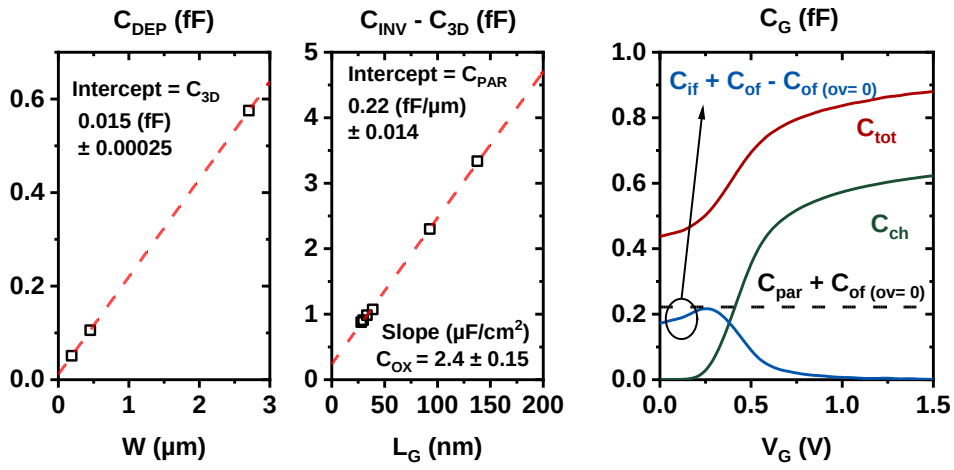


Figure 2-7(Left) Variation of depletion capacitance over width for devices with gate length of 30nm. The interception gives the 3D parasitic capacitance that is not dependent on the device length. (Middle) Variation of inversion capacitance over device length. The slope gives the value of C_{ox} while the intercept correspond to C_{2D}. (Right) Gate capacitance components of a 27nm MOSFET. The channel capacitance increases from zero to C_{ox} while the access capacitance decrees to zero in strong inversion.

The extraction of the parasitic capacitances from the linear fit of C_{inv} & C_{dep} is an important step in the complete process of the junction profile extraction. A large range of device lengths must be used for a better accuracy witch can become unmanageable. Nonetheless, the parasitic extraction can be uncomplicated by using two large devices (W > 1µm) of short and long (L_G >1µm) gate lengths. For the long devices, the fringe and parasitic capacitance components are much smaller than the oxide capacitance and can be considered as negligible. Similarly, the 3D component

is negligible on both short and large devices. Therefore, the C_a component is obtained from only two measurements using eq. 2-4.

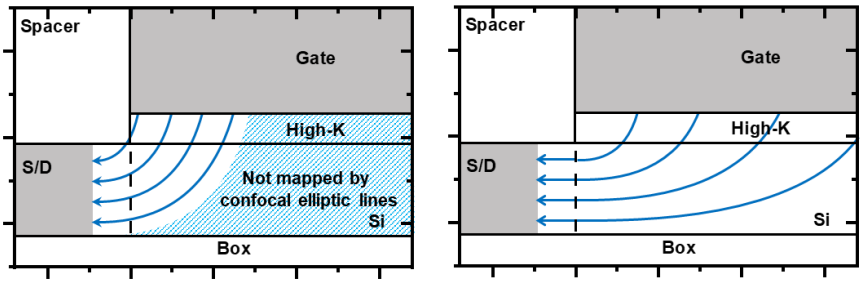
$$C_{ga}^{w} = \frac{C_{dep\_short} - \left(C_{inv\_sort} - \frac{L_{G\_short}}{L_{G\_long}} C_{inv\_long}\right)}{2W} \quad (2\text{-}4)$$

It is also important to note that this description considers a symmetrical device. However, the transistor can sometimes be asymmetrical. This is the case when the source is internally connected to the substrate or when the source sees a different implant condition than the Drain. For those cases, the devices must be measured between the gate and only one of the access and the factor 2 on the previous equations must be adjusted accordingly.

2.2.2 Depletion capacitance modelling based on conformal mapping

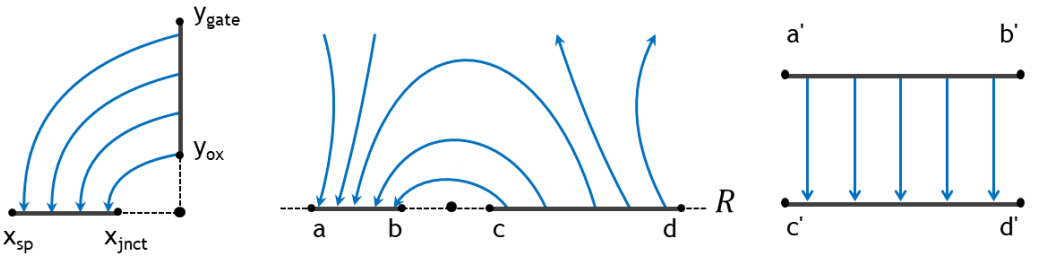
To obtain a precise but simple model for the fringe capacitance between gate and access we need to separate it in elementary components. Each element can then be approximated by parallel or perpendicular plate capacitors. **Erreur ! Source du renvoi introuvable.** presents the three capacitive components for the MOSfet that are dependent on the junction position. For complete modelling of the device whether its junction position, it is necessary to consider the overlap capacitance (C_{ov}) in addition of C_{of} and C_{if}. This parasitic component exists only when the junction overlaps the gate. It corresponds to the plate capacitor between the overlap region in the channel and the gate oxide (see Fig 2.6). Other parasitic capacitances, like the coupling between gate and epitaxial regions and interconnections are not dependent on the junction position. Those fixed components have been subtracted from raw measurements using the former pre-step i.e. the interception at zero from the capacitance at strong inversion over gate length.

Previous propositions of using conformal mapping to model the fringe capacitance in the transistor used confocal elliptical lines to map the electrical field and directly obtain the expression of capacitance without the need to solve the entire potential [refs models]. However, this method has a drawback: the assumption of confocal elliptic lines is not valid for a significant portion of the potential map as illustrated in Figure X. As a result, this solution overlooks a crucial component of the inner and outer fringe capacitance and therefore necessitates the inclusion of several additional terms to attain good accuracy.



Sketch illustrating the contrast between the classic field map that employs confocal elliptic lines (on the left) and the model that is not restricted to the same focal point for the ellipses (on the right).

The analytic expressions for each capacitance component are then obtained using the more accurate technique proposed by [ref]. In this paper, an approximation for the fringe capacitance $C_{1\to2}$ between two electrodes is obtained using two mathematical conformal transformation as illustrated in figure X.



The first transformation maps the field onto imaginary plane where the electrodes are placed at the Real axis. 7At the example of the two perpendicular plates from figure X, the analytic function ($z \rightarrow z^2$) directly maps the structure (a) to (b). Then it uses the Schwarz–Christoffel theorem as proposed by [ref] to obtain a second conformal mapping function transforming the structure from (b) to a parallel electrodes capacitance (c), as shown in Fig. 2. The capacitance is then calculated using Eq. X where $K_{\mathrm{el}}(K)$ is the complete elliptic integral, with K its modulus. The expression of K can be calculated for outer and inner components based on the parameters (a, b, c & d) that are derived from the devices geometries.

$$C(K) = \epsilon \frac{K_{\mathrm{el}}(K)}{K_{\mathrm{el}}\left(\sqrt{1-K^2}\right)} \; where \; K = \sqrt{\left(\frac{b-a}{c-a}\right)\left(\frac{d-c}{d-b}\right)}$$

Using the example of two perpendicular plates presented at figure X, the modulus K is given by eq. X after following the $(z \rightarrow z^2)$ transformation. Nonetheless, in order to avoid numerically integrating every elliptical integral, the expression 2-4 [ref] is used as a very useful mathematical approximation for eq. X.

$$K = \sqrt{\left(\frac{{x_{sp}}^2 - {x_{jnct}}^2}{{x_{sp}}^2 + {y_{ox}}^2}\right)\left(\frac{{y_{gate}}^2 - {y_{ox}}^2}{{y_{gate}}^2 + {x_{jnct}}^2}\right)}$$

$$C(K) = \begin{cases} W\frac{2\varepsilon_r}{\pi} \cdot \ln\left(2\sqrt{\frac{1+K}{1-\mathrm{K}}}\right), & \mathrm{K} \geq \frac{1}{\sqrt{2}} \\ W\frac{2\varepsilon_\mathrm{r}}{\pi} \cdot \ln\left(2\sqrt{\frac{1+\sqrt{1-K^2}}{1-\sqrt{1-K^2}}}\right)^{-1}, & \mathrm{K} \geq \frac{1}{\sqrt{2}} \end{cases} \quad (2\text{-}4)$$

In [ref], the proposed model is applied to the case of MOSFETs but does not take in account the case of underlap devices. In this thesis, a modified version is proposed suitable for any overlap position. The modified expressions are given below:

2.2.2.1 Inner Fringe Capacitance Cif

For the internal fringe capacitance, [ref] proposes the expression 2-5 for $\mathrm{K_{if}}$ that takes in account the total surface of the gate and uses a mirror boundary instead of simply dividing it on two halves. The value of $\mathrm{EOT_{si}}$ corresponds to the effective silicon thickness of the gate oxide i.e. considering the permittivity of crystalline silicon of 11.9. $\mathrm{T_{si}}$ is the silicon film thickness and ov corresponds to the value of the junction overlap with respect to the gate. The variable ov is positive or negative for overlapped and underlapped devices respectively: When ov equals to zero, the junction is positioned at the interface between gate and spacer.

$$\mathrm{K_{if}} = \frac{1}{\cosh\left(\pi\frac{\mathrm{EOT_{si}}}{\mathrm{L_g} - 2\cdot ov}\right)} \frac{\sqrt{1 - \left[\frac{\cosh\left(\pi\frac{\mathrm{EOT_{si}}}{\mathrm{L_g} - 2\cdot ov}\right)}{\cosh\left[\pi\frac{\mathrm{EOT_{si}} + \mathrm{T_{si}}}{\mathrm{L_g} - 2\cdot ov}\right]}\right]^2}}{\sqrt{1 - \left[\frac{1}{\cosh\left[\pi\frac{\mathrm{EOT_{si}} + \mathrm{T_{si}}}{\mathrm{L_g} - 2\cdot ov}\right]}\right]^2}} \quad (2\text{-}5)$$

$$where: \mathrm{EOT_{si}} = \mathrm{EOT}\frac{\varepsilon_{si}}{\varepsilon_{ox}}\ and: \mathrm{EOT} = \frac{\varepsilon_{ox}}{\mathrm{C_{ox}}}$$

The modified expression of the inner fringe capacitance C_{if}, that accounts for both overlapped or underlapped junctions, is given in equation 2-6. When the device is overlapped, the capacitance is directly given by the equation 2-4. However, for underlapped devices, the capacitance is calculated from the association in series of the inner fringe capacitance at $ov = 0$ with a parallel plate capacitance representing the underlapped region.

$$C_{if}^{w} = \begin{cases} k_{box} \cdot C\left(\mathrm{K_{if}}(ov)\right), & \mathrm{ov} \geq 0 \\ k_{Box} \cdot \dfrac{C\left(\mathrm{K_{if}}(0)\right) \cdot \varepsilon_{si}\dfrac{\mathrm{T_{si}}}{ov}}{C\left(\mathrm{K_{if}}(0)\right) + \varepsilon_{si}\dfrac{\mathrm{T_{si}}}{ov}}, & \mathrm{ov} < 0 \end{cases} \quad (2\text{-}6)$$

$$where: k_{Box} = 1 - \frac{\mathrm{EOT}}{T_{Box}}$$

2.2.2.2 Outer Fringe Capacitance Cof

The expression of K_{of} for the outer fringe capacitance is given by 2-7 where $\mathrm{T_{ox}}$ is the physical oxide thickness, $\mathrm{L_{sp}}$ is the spacer lateral thickness and $\mathrm{H_{epi}}$ is the height of the access epitaxy, as depicted in Fig 2.6. This equation differs from the simple perpendicular plates example as it also takes in account the lateral part of the raised source and drain [ref]. The components of the outer capacitance due to the remainder of the epitaxial regions and interconnections are not considered in this model because we are only interested in the components that vary with the overlap position. An exact knowledge of the outer fringe capacitance, including these components, is not mandatory for the proposed method.

$$\mathrm{K}_{of} = \sqrt{\frac{\cos\left(-\frac{\pi \cdot ov}{\mathrm{L_{sp}}}\right) + \cosh\left(\frac{\pi \cdot \mathrm{H_{epi}}}{\mathrm{L_{sp}}}\right)}{\cosh\left(\frac{\pi \cdot \mathrm{T_{ox}}}{\mathrm{L_{sp}}}\right) + \cosh\left(\frac{\pi \cdot \mathrm{H_{epi}}}{\mathrm{L_{sp}}}\right)}} \cdot \sqrt{\frac{\cosh\left(\frac{\pi \cdot \mathrm{H_{epi}}}{\mathrm{L_{sp}}}\right) - \cosh\left(\frac{\pi \cdot \mathrm{T_{ox}}}{\mathrm{L_{sp}}}\right)}{\cosh\left(\frac{\pi \cdot \mathrm{H_{epi}}}{\mathrm{L_{sp}}}\right) - \cos\left(-\frac{\pi \cdot ov}{\mathrm{L_{sp}}}\right)}} \quad (2\text{-}7)$$

In other to correctly describe the measurements, the expression of outer fringe capacitance takes in account the subtraction of the capacitance components from the first step. For overlapped devices, the variation of outer fringe capacitances between depletion and strong inversion equals to zero because it is the same as the reference. For devices that are underlapped, the outer fringe capacitance is smaller in depletion

compared to strong inversion. We developed, therefore, the expression 2-8 where $\Delta C_{of} = 0$ for overlapped devices and $\Delta C_{of} < 0$ for underlapped devices.

$$\Delta C_{of}^{w} = \begin{cases} 0, & \text{ov} \geq 0 \\ C\left(\mathrm{K_{of}}(ov)\right) - C\left(\mathrm{K_{of}}(0)\right), & \text{ov} < 0 \end{cases} \tag{2-8}$$

2.2.2.3 Overlap Capacitance Cov

The overlap capacitance is simply given by the parallel plate expression x when the devices is overlapped. Evidently, for underlapped devices this capacitance is equals to zero.

$$C_{ov}^{w} = \begin{cases} \varepsilon_{ox} \dfrac{\text{ov}}{EOT}, & \text{ov} \geq 0 \\ 0, & \text{ov} < 0 \end{cases} \tag{2-9}$$

Finally, the total access depletion capacitance is given by the sum of the three components

$$C_{ga}^{w} = C_{if}^{w} + C_{of}^{w} + C_{ov}^{w} \tag{2-10}$$

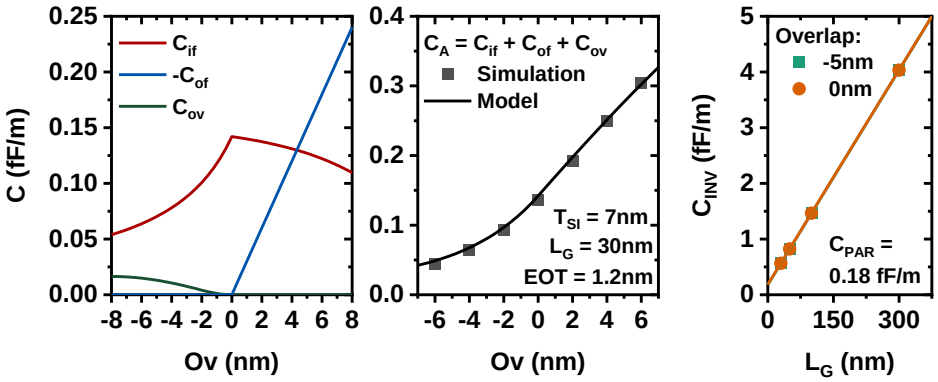


Figure 2-8 Validation of the analytical model for access capacitance C_A versus overlap position for a 30nm device. (Left) Contributions of each parasitic capacitance to C_A . (Center) Total access capacitance obtained from the model (C_{ga}) against TCAD simulation. (Right) Extraction of parasitic capacitance form simulation at different overlap conditions.

In order to verify the model, we performed TCAD electrostatic simulations of FDSOI devices with different dimensions and overlap positions. To eliminate uncertainties related to variations in overlap resulting from the depletion of access, the junctions were selected to be metallic and abrupt for the simulation. **Erreur ! Source du renvoi introuvable.** (left) presents the variation of the different components (C_{if}, C_{of} *and* C_{ov}) derived from the aforementioned analytical model. Regarding C_{if}, it reaches a maximum for a zero overlap. For underlapped devices, the capacitance

reduces as the distance between the junction and the gate increases. For overlapped devices, C_{if} also reduces because of the decreases of the gate surface. Seemingly, C_{ov} is zero for underlapped devices and increases linearly with overlap (ov>0).

The parasitic capacitance is extracted from the simulation results using the same method from the previous section. **Erreur ! Source du renvoi introuvable.** (right) presents the variation of C_{inv} for two simulated devices with $ov = -5nm$ and $ov = 0nm$ as a function of the device length. As stated before, independent of the value of overlap, the parasitic capacitance extracted from strong inversion regime is the same because it is referenced to an overlap of zero and the access region is completely inverted on both cases.

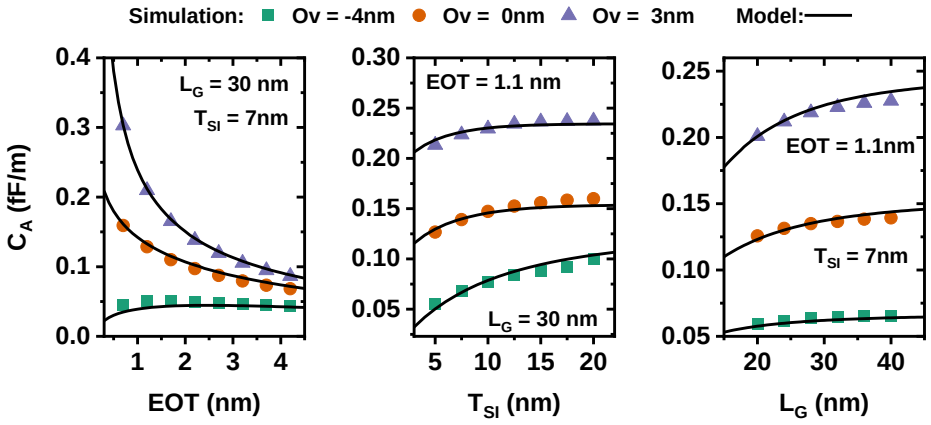


Figure 2-9 Validation of C_A model with TCAD simulation for different devices geometry at three overlap conditions ($ov = -4, 0\ and\ 3\ nm$). From left to right: Variation of Equivalent oxide thickness, silicon channel thickness and gate length.

In addition, we checked the validity of the model against TCAD simulations of devices with different dimensions. **Erreur ! Source du renvoi introuvable.** compares the results of simulation and model over the silicon film thickness, the oxide thickness and the channel length. The model correctly reproduces all simulation results for each overlap condition. It is clear that the C_A is strongly dependent of the EOT value and therefore must be correctly extracted from measurements. Furthermore, the variation of C_A with overlap increases for thinner gate oxides. This means that the model becomes less sensitive to measurement noise and errors on the parasitic components extraction when applied to devices with EOT < 2nm.

2.2.3 Method CV-AJP to extract device junction profiles

The CV-AJP method is illustrated in **Erreur ! Source du renvoi introuvable.**. It consists of two steps: (1) Extracting the carrier junction position as a function of gate bias using measurements of the MOSFET capacitance in the depletion regime (2) Extracting of the density of dopants for each overlap position.

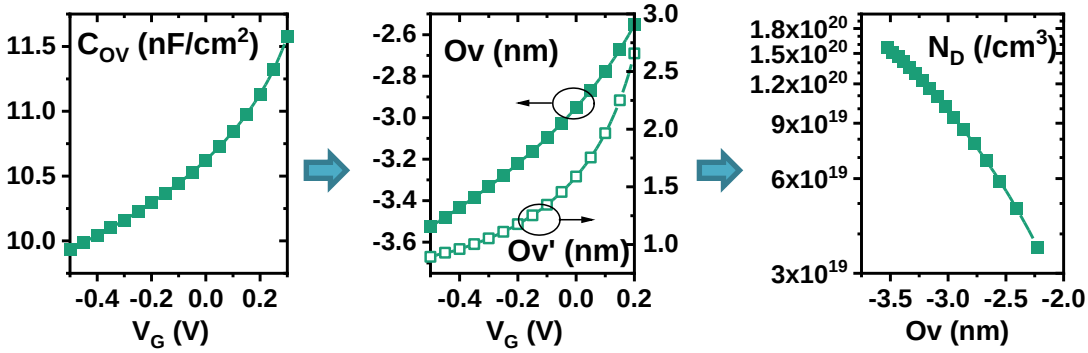


Figure 2-10 (Left) Measured capacitance in depletion regime. (Middle) Extraction of overlap position and its derivative at each V_G bias from conformal mapping model. (Right) Extraction of doping profile from equation X.

More precisely, the first step consists in extracting the overlap variation as a function of V_G. To do that, we use the unique relation that relates the access capacitance to overlap for $V_G < V_T$ and a given device geometry. C_{ga} is directly extracted from device capacitance measured according to eq. 2-3 or 2-4, and well modelled by the previous conformal mapping model (eq. 2-10).

For the second step, we obtain the dopant concentration as the function of the variation of $Ov(V_G)$ using eq. 2-12. We consider here that, in depletion regime, the variation of charges responsible for the capacitance is only ascribed to carriers in the access region and not in the channel (eq. 2-11). Therefore, by knowing the variation of the depletion volume ($dO_V \cdot W \cdot T_{si}$) and the variation of charges (C_{ga}) we can derive the number of dopants that are depleted during the interval dV_G.

$$\frac{dQ_{Ov}}{dV_G} = C_{ga}(V_G) \;\; for \;\; C_{Ch-SD} \gg C_{Inv} \tag{2-11}$$

$$N_{dop} = \frac{C_{ga}(V_G)}{q \cdot \frac{dO_V}{dV_G} \cdot W \cdot T_{si}} \tag{2-12}$$

Finally, we can combine the position of the depletion region from step 1 and the number of depleted dopants from step 2 to obtain the access profile. **Erreur ! Source**

du renvoi introuvable. highlights the suitability of the technique to extract the dopant profile (N_D) in FD-SOI devices with a silicon thickness of 6nm. On the left, we can see the capacitance obtained by simulating devices with different profiles. On the right, we see that the profile obtained using the CV-AJP method with this capacitance curves perfectly reproduces the Gaussian profiles entered as inputs in TCAD simulations.

Nonetheless, the CV-AJP is not applicable to any devices. Since it approximates the volume of depletion of the junction in the interval of dV_G as a cuboid with constant doping density, it requires that the following conditions are true:

A. The doping profile is much stronger along the horizontal direction (channel) than along the vertical direction (T_{si}). In other words, the dopants of source and drain are considered as constant along the silicon film thickness. This statement holds particularly true when T_{si} is small.

The film is thin enough, so that the depletion of majority carriers in the access occurs mainly along a single horizontal axis from gate to access.

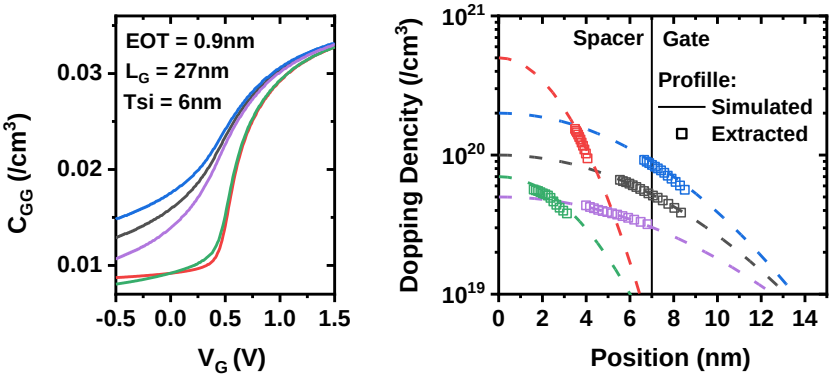


Figure 2-11 Demonstration of technique with TCAD simulations. (Left) Simulated capacitance for different Gaussian profiles. (Right) Extraction of doping profile for each case compared to the Gaussian profile used for each simulation.

It is clear that conditions A and B are met only for advanced devices that contain ultrathin silicon thickness (<10nm). It is important to note that the method extract only the density of electrically active dopants that are incorporated into the lattice.

2.2.4 Method validation on real device

The technique is validated against the capacitance measurements of UT-FDSOI devices. **Erreur ! Source du renvoi introuvable.**-left presents the extracted profile for two devices with different lengths. Even though the raw capacitance measurements are different between them, the extracted profiles are almost the same. This is possible

thanks to an accurate modelling of the variation of the fringe capacitance with device length **Erreur ! Source du renvoi introuvable.**-right.

To do this extraction, the parasitic capacitances $C_{2D} + C_{3D}$ has been formerly assessed using the technique highlighted in the previous section. **Erreur ! Source du renvoi introuvable.**-middle reports the sensitivity of the technique to this parasitic capacitance evaluation, the most sensible parameter of the method. A 10% error in C_{PAR} leads to a 1nm variation in the overlap position, but hardly affects the overall shape of the profile. This demonstrates the robustness of the CV-AJP method for the quantitative assessment of the true concentration of electrical active dopants.

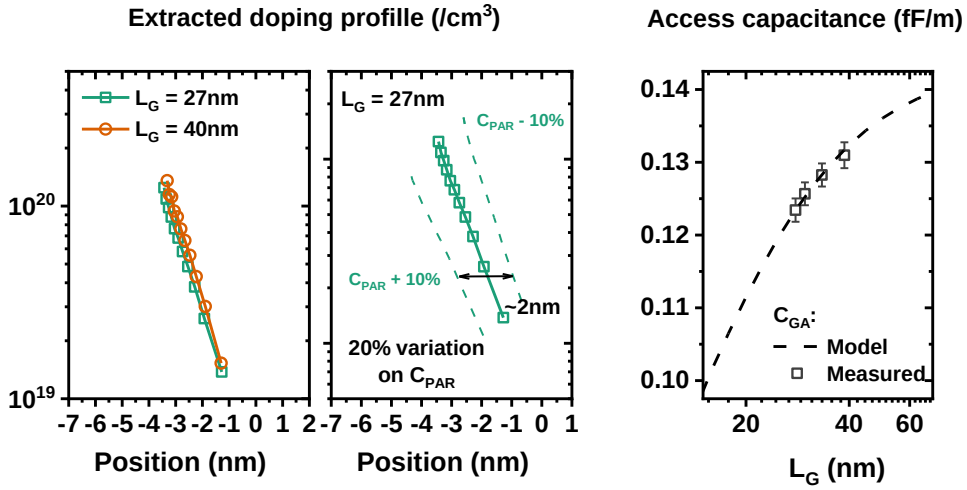


Figure 2-12 CV-AJP applied to with UT-FDSOI devices with $T_{SI} = 6nm$ and $EOT = 1.35nm$. (Left) Profile extracted from devices with gate length of 27nm and 40nm. (Middle) Variability of extracted profile considering an uncertainty on the extraction of parasitic capacitance of +/- 10%. (Right) Measured vs model of C_A for devices with different gate lengths.

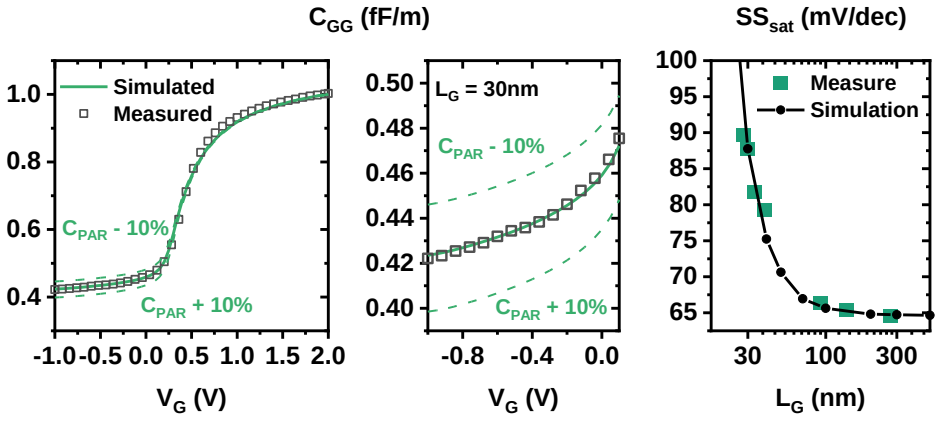


Figure 2-13 Validation of extracted profile with TCAD simulation for a device with $L_G = 30nm$. (Left and Middle) Comparison of Gate capacitance measurement and TCAD simulation with the profile extracted by the CV-AJP method. The simulations with an +/- 10% incertitude on C_{PAR} are also plotted. (Right) Subthreshold slope extracted from TCAD simulation with the extracted profile compared to the measurement.

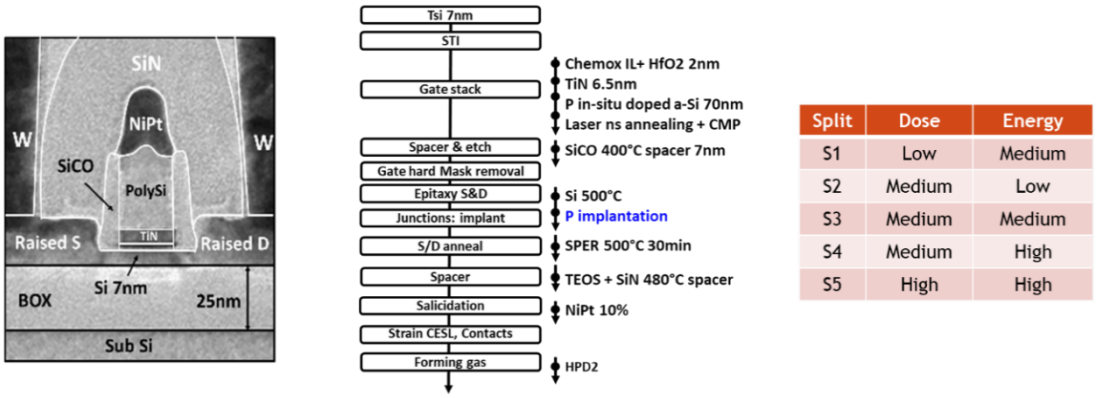
The effectiveness of the method is highlighted in **Erreur ! Source du renvoi introuvable.**. The post TCAD simulations of the device CV characteristics using the extracted profile from our CV-AJP technique perfectly reproduces the raw capacitance measurement. In addition, a good match is achieved between simulations and measurements for the subthreshold slope as well. It can be noted that this method can be a powerful tool to calibrate TCAD simulations for the device.

2.3 STUDY OF LTB DEVICES

On this session, the CV-AJP technique is used to evaluate the performances and activation level that can be obtained on low thermal budget devices activated using a SPER anneal at 500°C

2.3.1 Phosphorous implant effects on NMOS devices

Fig. X-middle shows the LTB (≤500°C) N-FDSOI device process flow fabricated with <100> oriented channel (T_{Si} = 7nm). The gate stack consists of a HFO2/TiN (EOT=0.91nm) with UV nanoseconds laser anneal (UV-NLA) for poly-Si activation. Low-k SiCO spacers ($\varepsilon r = 4.5$), Si Raised Source Drain (RSD) epitaxy followed by SPER activation @ 500°C are carried out. Various phosphorous implantations conditions in energy and dose (S1-S5) have been studied (**Erreur ! Source du renvoi introuvable.**-right). In this process, High Pressure Deuterium (HPD2) and a tensile CESL are also included for improved Dit passivation and mobility boost respectively.



Split	Dose	Energy
S1	Low	Medium
S2	Medium	Low
S3	Medium	Medium
S4	Medium	High
S5	High	High

Figure 2-14(Left) cross-section HRTEM of a device with 27nm channel length and 7nm silicon film thickness. (Middle) LTB N-FDSOI device process flow. (Right) Five split variants for phosphorous implantation.

Erreur ! Source du renvoi introuvable. shows the CV capacitances measured for four (S1-S4) variants and the corresponding active donor profiles extracted with our CV-AJP technique. As expected, increasing the energy or the dose allow the dopants to diffuse closer to the gate while reducing the abruptness of the junction. The technique is validated against TCAD simulations for all the splits. A perfect matching between the measurements and the simulations is highlighted, confirming once again the suitability of the technique. Additionally, as shown in Figure 2 16 (left), the A/C interface is anticipated to be located far away from the junction position.

Consequently, the profiles in the figure correspond to a thermal activation at 500°C. This a differently from what is expected from the solubility limit found in the literature with a low temperature anneal is sufficient to activate the dopants under the spacers.

One explanation for the activation on low temperature is that the value of solubility obtained from the literature is often given for impurities on bulk silicon. For UT-FDSOI devices, the dopants are very close to two oxide interfaces and the solubility may be greater than expected. Oxide interfaces behave like interstitial sinks, meaning that it tends to absorb interstitial defects introduced by the impurities implantation [ref]. Indeed, a lower density of interstitials near the spacer and BOX interface would reduce the probability of formation of clusters and increase the dopant solubility near the region [ref]. Another possibility is that although not fully amorphous, vacancies generated during the implantation process are "healed" during the low temperature annealing, which could result in the incorporation of dopants.

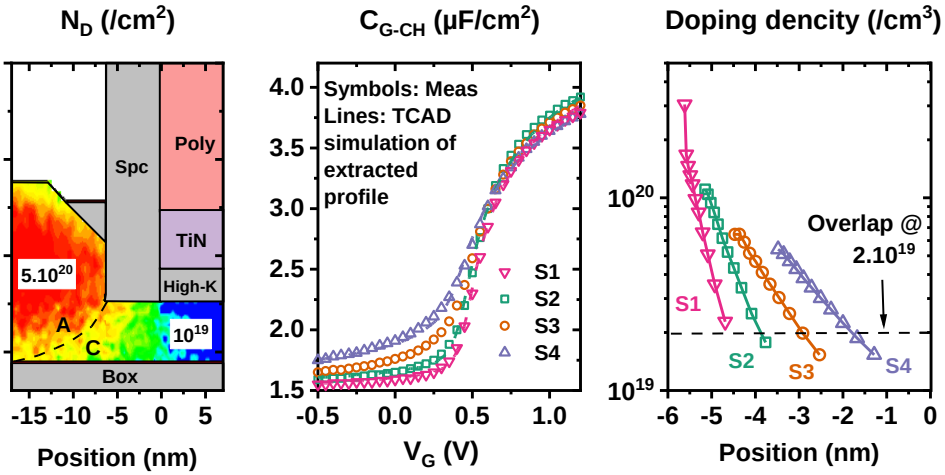


Figure 2-15(Left) KMC simulation of the implant process illustrating the expected dopants profile and the position of the amorphous–crystalline interface in the access region of the devices. (Middle) Measured gate capacitance curves of the devices from four (S1-S4) variants compared to the TCAD simulation using each extracted profile. (Right) Profiles extracted from variants S1 to S4.

The technique is then used to analyse the device performance and to draw guidelines to improve it. **Erreur ! Source du renvoi introuvable.** presents classic IV FoM for each implant condition. As expected, by extending the junction over the spacer (reduced underlap region) and reducing the abruptness of the junction, SCE becomes stronger. Indeed DIBL, SS increases with overlap while $V_T(L)$ decreases in magnitude for shorter channel lengths. For S1, an abnormal behaviour is yet observed. Due to the small amount of dopants near the interface with the spacer oxide, this device

becomes highly sensitive to the repealing effect produced by a small amount negative charges in the SiCO spacer [5]. These negative charges deplete the access region and increase the access resistance of the device.

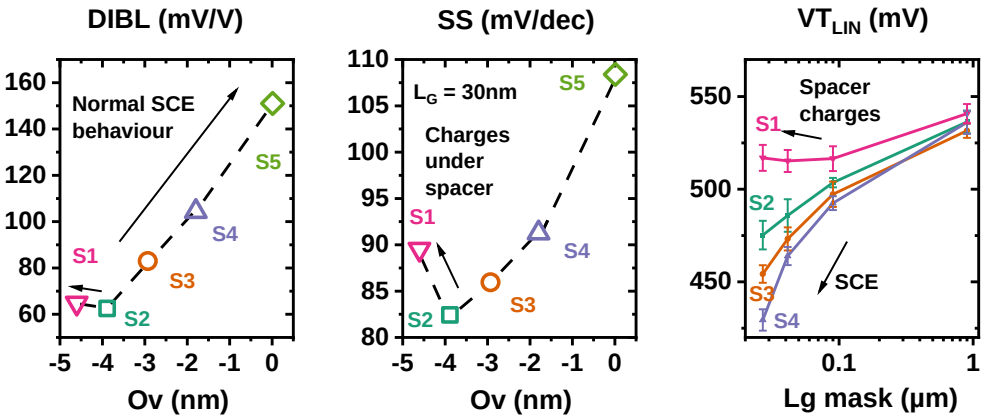


Figure 2-16 DIBL (Left) and subthreshold slope (Middle) of a 27nm device extracted for each implantation variant (S1-S5). (Right) variation of linear threshold voltage over channel length obtained from each variant.

To understand the impact of the dopant profile on the access resistance, R_{acc} is extracted using the modified Y function method [6] and plotted as the function of the junction position for 4 splits (Fig X). This method extracts two components for the access resistance, a fixed contribution R_0 that is independent on the gate bias and the parameter σ that accounts for the variation of access resistance with gate bias, as the carrier accumulation in the access at strong inversion conditions. The method requires the extraction of the slope and the intercept from the linear fitting of parameters θ_1 and θ_2 versus β as illustrated at **Erreur ! Source du renvoi introuvable.**.

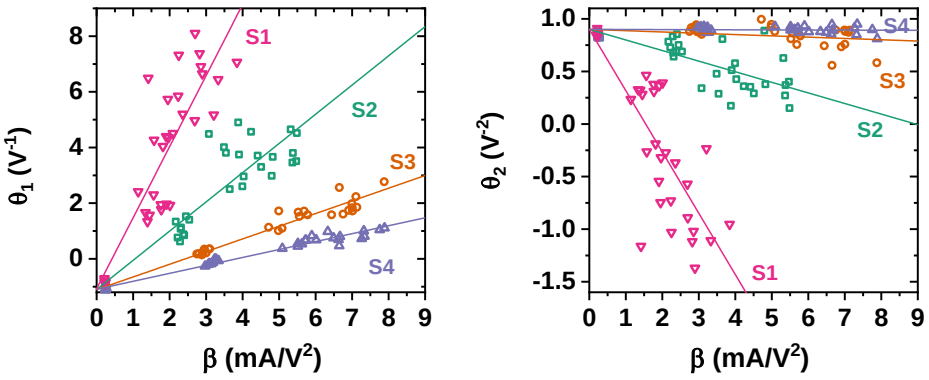


Figure 2-17 Extraction of parameters for modified Y function method. From slope and intercept of the linear interpolation of θ_1 and θ_2 over , we are able to extract access resistance parameters R_0 and σ. [ref]

The benefit of reducing the device underlapping on R_{acc}, by optimizing the phosphorous implant, is twofold. Firstly, a larger number of dopants strongly decreases the access resistivity and therefore the R_0 component. Secondly, it reduces the sensitivity of this resistance to gate bias, resulting on a reduced electrostatic control of the gate over the spacer region (**Erreur ! Source du renvoi introuvable.**).

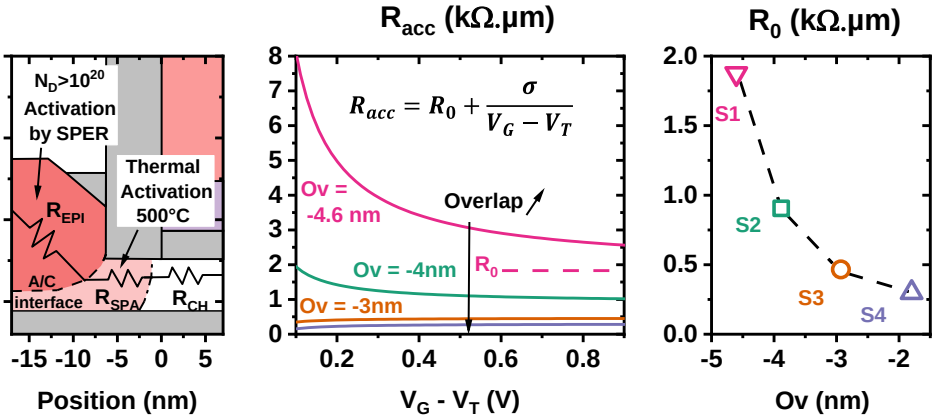


Figure 2-18 (Left) sketch illustrating the different resistive components that are dependent on the activation level and profile in the access. (Middle) Access resistance extracted from Y function for each implantation variant (S1-S4). (Right) Fixed component R_0 extracted for each variant versus overlap.

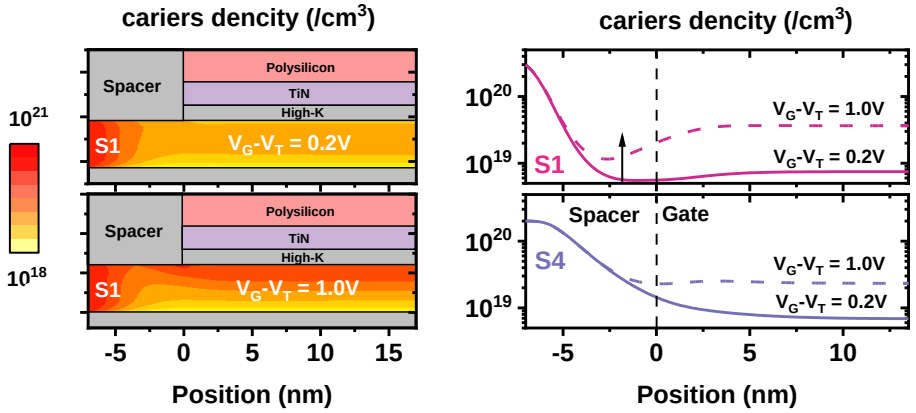


Figure 2-19 (Left) 2D Negative carrier density obtained from TCAD simulation of variant S1 at two gate bias conditions ($V_G - V_T = 0.2V$ and $V_G - V_T = 1.0V$). (Right) 1D cutline at the middle of the channel of electron density for variants S1 and S4 at both bias conditions.

Erreur ! Source du renvoi introuvable. illustrates the effect of the dependence of access resistance with gate bias for underlapped devices. Fig. X-left presents the density of carriers in the channel simulated using the S1 profile. We can see that the carrier density under the spacers increases when overdrive increases from $V_G - V_T = 0.2V$ to $V_G - V_T = 1.0V$ thereby reducing the access resistance. **Erreur ! Source du renvoi introuvable.**-right, compares the carrier densities for both overdrive conditions

of variants S1 and S4. In the case of a more overlapped device, where the majority of carriers originate directly from dopants, the variation of access resistance induced by the gate bias is reduced.

Finally, **Erreur ! Source du renvoi introuvable.** presents the I_{ON} and I_{OFF} FOMs plotted as the function of the overlap position. The key feature is that both follows a single trend with overlap. I_{OFF} exponentially increases as SCE becomes important, while I_{ON} rises almost linearly mainly due to the combined effect of R_{ACC} and Vt_{SAT} reduction. From this picture, we could predict that reducing the overlap to the position of variant S5 could lead to $I_{OFF} = 10^{8}\ A/\mu m$ & $I_{ON} = 900\mu A/\mu m$ for this LT technology. This could be done on this case by further increasing energy and dose of the phosphorous implant. However, this solution has been attempted here in S5 but, unfortunately, has not been successful. The main reason for that was the partial recrystallization of the Si in the access region since an almost vertical amorphous/crystalline interface below the spacer cannot properly acts as a seed for a complete crystalline regrowth at 500°C with 30 minutes anneal [7].

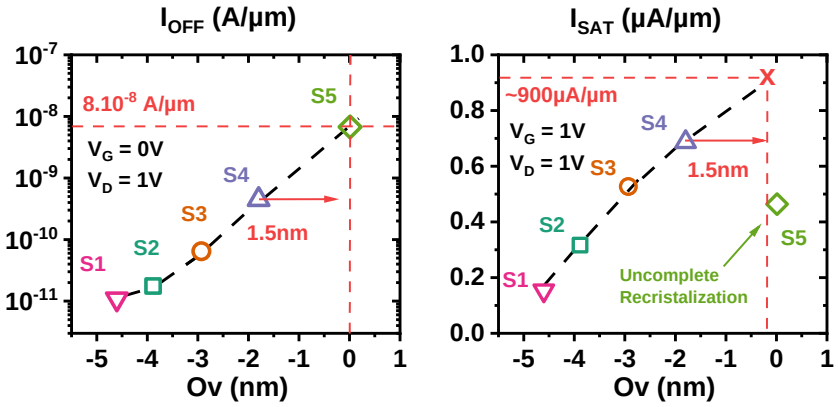


Figure 2-20(Left) Off-state current ($V_G = 0V$ and $V_D = 1V$) and (Right) On-state current ($V_G = 1V$ and $V_D = 1V$) plotted against the metallurgic overlap position from each variant (S1-S5).

Another potential solution to reach the overlap of the variant S5 may be to reduce the thickness of the spacer of 1.5nm. With the reduction of the spacer thickness on a extension-last integration scheme, the transistor is expected to have an I_{ON} near 900µA/µm at I_{OFF} equal to only $8\ nA/\mu m$. This would be a further improvement of the device performance, considering that other FoMs like DIBL and SS remains low. Nevertheless, the optimization made already here with the S4 variant leads to $I_{ON} =$

$940\mu A/\mu m$ @ $I_{OFF} = 100nA/\mu m$ that already outperforms literature results for LT technology [8].

2.3.2 Effects of etching of the sacrificial oxide on NMOS devices

After the epitaxy of source and drain and before the phosphorous implantation, it is necessary to remove the SiN hard mask used to protect the gate. However, this process can also partially etch the spacer oxide (SiN or SiCO). Therefore, an oxide layer is deposited to protect the spacers during the removal of the hard mask. This oxide tends to fill the faceted region of epitaxies, which reduces the implantation distance at the access region under the spacers. In order to improve the overlap position without changing the implant conditions, we tried to etch the sacrificial oxide before the phosphorous implant.

The variants S2d and S3d are presented on **Erreur ! Source du renvoi introuvable.**-left. Both have the same phosphorous implant conditions of variants S2 and S3. The only distinction is the oxide-etching step that is performed before the implant. **Erreur ! Source du renvoi introuvable.**-right show the KMC simulation results for variants S3 and S3b. By removing the oxide layer, we can see that the dopants are more deeply implanted into the channel, leading to a more overlapped device.

Split	Dose	Energy	Nat. oxide
S2	Medium	Low	Present
S3	Medium	Medium	Present
S2d	Medium	Low	Etched
S3d	Medium	Medium	Etched

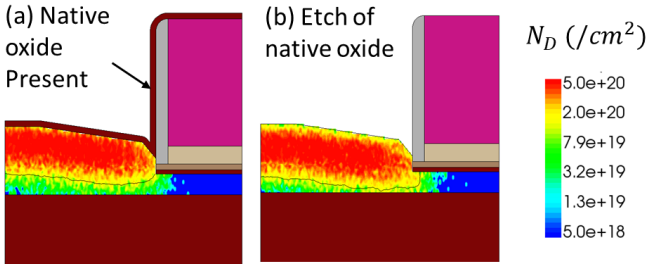


Figure 2-21(Left) Table presenting the 4 split variants. S2 and S3 are equivalent to variants of the previous session, S2d and S3d but with the etching of the oxide above the source and drain. (Right) KMC simulation presenting the position of dopants for variants S3 and S3b.

The CV-AJP technique is applied on each variant and the extracted profile is shown at **Erreur ! Source du renvoi introuvable.**-left. As expected, there is an increase on the overlap position of around 1nm for etched variants. Additionally, there is a remarkable similarity between the profiles extracted for variants S3 and S2d, indicating that the oxide etching is equivalent to an increase of 1keV in implant energy. Figure X-middle presents the θ_1 over β plot for each variant indicating that the access

resistance actually drops for the etched variants. This is due to the fact that the removal of the oxide increases the devices overlap and thereby reduces access resistance. On the other hand, **Erreur ! Source du renvoi introuvable.**-right shows that an increase in energy implant has a greater effect on access resistance than etching. It is possible that more defects will be created on the junction as the amorphization depth increases.

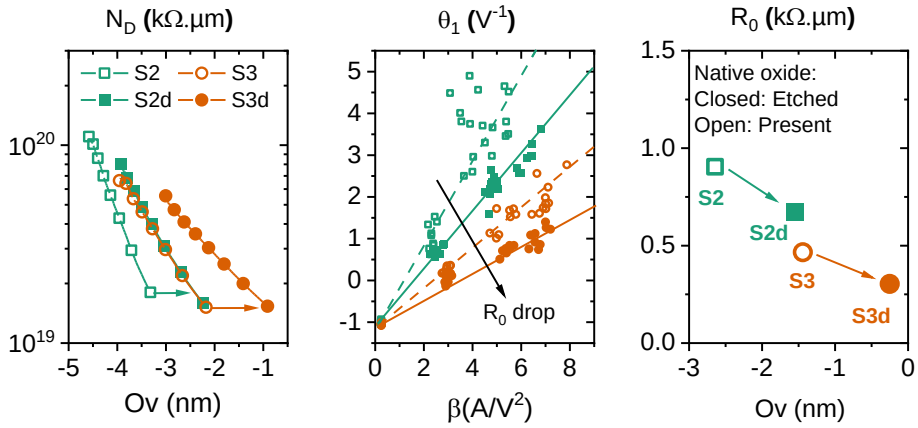


Figure 2-22(Left) Extraction of density of dopants using CV-AJP technique for each variant. (Middle) Extraction of θ_1 over β from parameters Y2 function for each variant. (Right) Fixed component R_0 from access resistance obtained from of each variant.

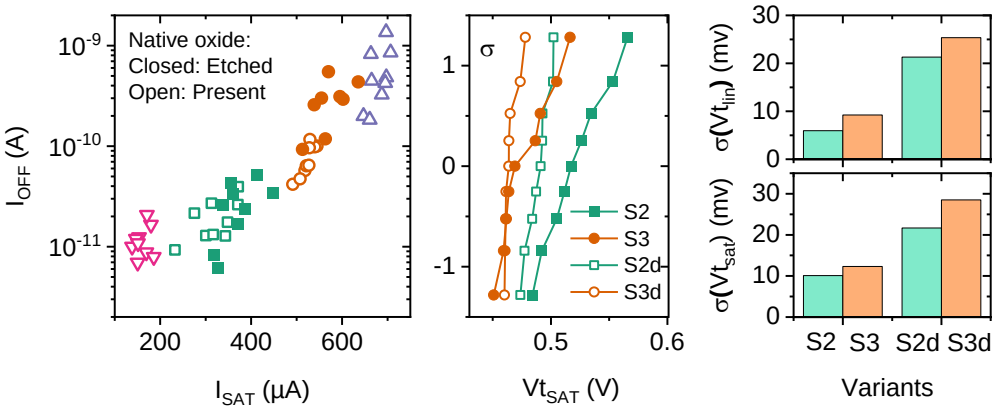


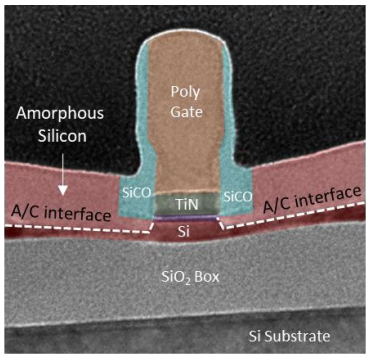
Figure 2-23(Left) I_{ON} vs I_{OFF} FoM for etched variants (S2d and S3d) compared to non-etched variants (S1 to S4). (Middle) Distribution of extracted values of saturation threshold voltage for each variants. (Right) Standard deviation of Vt_{LIN} and Vt_{SAT} extracted for each variant

A comparison of the I_{ON} vs I_{OFF} characteristics of etched variants with those of previous non-etched variants is presented in **Erreur ! Source du renvoi introuvable.**-left. We can see that the saturation current is reduced for variants S2d and S3d at same off state current. In addition, the variability of device parameters increases for etched devices. This effect is further illustrated at **Erreur ! Source du renvoi introuvable.**-middle and right. The standard deviation of Vt_{LIN} and Vt_{SAT} increase by a factor of

two after the etch. Actually, the enhanced variability of the access resistance and the Vt at a same overlap value may results from the non-uniformity of the etching process. Therefore we can conclude that it is better to increase the doping energy to achieve a better overlap than to etch the oxide in the epitaxy cavity.

2.3.3 Effect of germanium amorphization on PMOS devices

With P-type devices, the amorphization step is carried out using Germanium implantation followed by doping bore implantation. **Erreur ! Source du renvoi introuvable.**-left shows the cross-section HRTEM of a LTB (≤500°C) P-FDSOI device after a high-energy (6keV) germanium implantation step. The figure highlights the amorphous-crystalline interface in the access region. For this implant condition, the region under the spacers also becomes amorphous. This indicates that the SPER activation process can also occurs near the junction. Due to the 2-step processes for amorphization and doping, a more deeper analysis on the correlation between amorphization and activation of bore dopants under the spacers has been performed. **Erreur ! Source du renvoi introuvable.**-right shows the bore and germanium variants used for this study.



Split	Ge Energy	B energy
S1A	Low	Low
S1B	Medium	Low
S1C	High	Low
S2A	Low	High
S2B	Medium	High
S2C	High	High

Figure 2-24(Left) cross-section HRTEM of a device with 27nm channel length and 7nm silicon film thickness after a High-energy Germanium implantation. The interface between amorphous and crystalline regions is clearly visible. (Middle) LTB N-FDSOI device process flow. (Right) Six split variants for Bore and Germanium implantation energy.

The CV-AJP technique is used again for each variant. The CV curves for variants S1A and S2A and the associated extracted profiles are shown in **Erreur ! Source du renvoi introuvable.**-left & **Erreur ! Source du renvoi introuvable.**-middle respectively. As expected, the strongest implantation energy results in an increase in the depletion capacitance and in overlap position. However, the junction profile stays the same independent of the germanium implantation condition as illustrated at

Erreur ! Source du renvoi introuvable.-right. This effect indicates that the maximum dopant activation level in this region does not depend on the recrystallization process. For the low energy germanium implantation variants, it is expected that a very low amount of dopants under the spacers would be activated by the SPER recrystallization. This is due to the low theoretical solubility of Bore on bulk silicon at around $10^{18}\ cm^{-3}$ [ref]. This implies that the thermal solubility near the spacer region should be greater than the density of dopants that were introduced in this region. This observation supports the hypothesis of increased impurity solubility near the two oxide interfaces, which may be attributed to the reduction of interstitials.

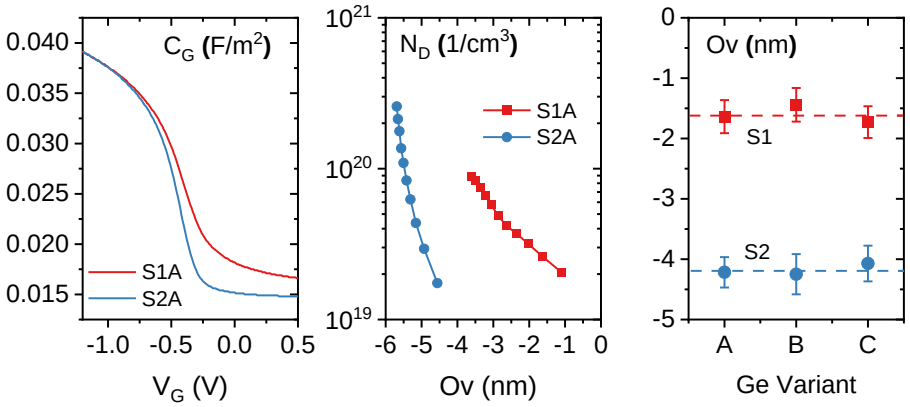


Figure 2-25(Left) Gate capacitance of a 27nm device from variants S1A and S2A. (Middle) Extraction of density of dopants using CV-AJP technique for each variants S1A and S2A. (Right) Overlap position extracted from CV-AJP plotted against the germanium implantation energy for both Bore implantation conditions.

The I_{ON} vs I_{OFF} FoM from variants S1A and S2A are presented on **Erreur ! Source du renvoi introuvable.**-left. As expected, increasing the overlap improves the performance of the device through the access resistance reduction. On the order hand, even if the junction position stays at the same position for all the variants, the saturation current reduces for high germanium energy implantation energies. This effect is more prominent for underlapped devices where the mean saturation current drops by a factor of two.

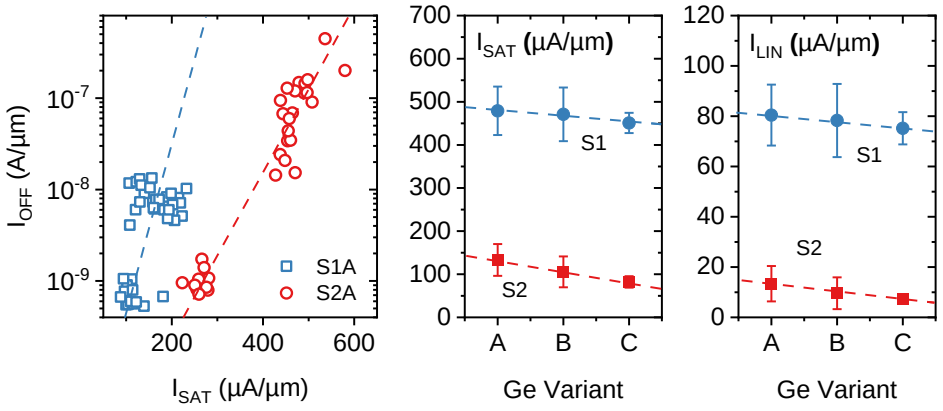


Figure 2-26(Left) I_{ON} vs I_{OFF} FoM from variants S1A and S2A (Middle) I_{SAT} and (Right) I_{LIN} plotted against the germanium implantation energy for both bore implantation conditions.

The performance degradation for the various Ge implantation splits is explained by a higher amount of defects near the junction. **Erreur ! Source du renvoi introuvable.** illustrates the effect of those traps on the threshold voltage. Unlike what is expected for short channel effects, the Vt_{LIN} increases for smaller devices. This effect is stronger for high Ge implantation energies. These traps are most probably related to end of range defects generated during the germanium amorphization. When the device is overlapped, the charge of ionized dopants is high enough to screen out these charged defects positioned under the spacers, thereby reducing the . This effect will be better explored in the next chapter. It is important to note that, even if the degradation on Vt is not seen on overlapped devices, the saturation and linear current still are impacted. Therefore we can conclude that the step of germanium implantation also plays a role on the access resistance of the devices.

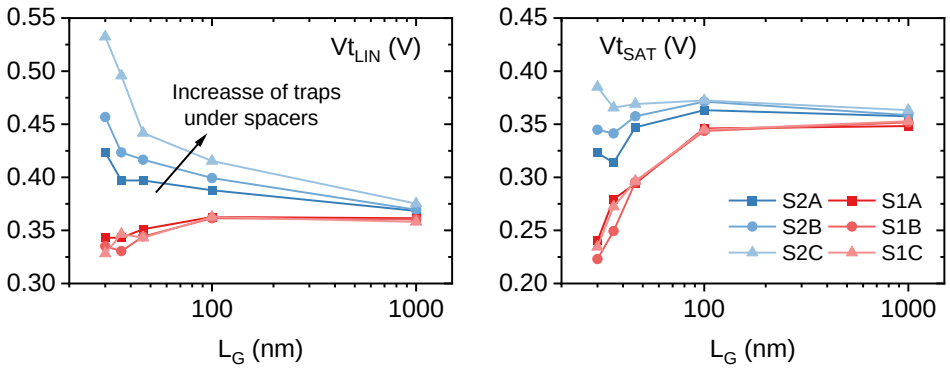


Figure 2-27(Left) Linear and (Right) Saturation threshold voltage plotted against channel length for each variant.

From this study, we can also conclude that is not necessary to perform a deep germanium implantation in order to activate dopants under the spacers with the

recrystallization method. On the contrary, it is better to perform a shallow amorphization that allows a high activation of dopants in the epitaxy region without modifying the junction shape and position.

2.3.4 X-first without amorphization

Considering that a pre amorphization step is not necessary to activate the dopants beneath the spacers, we can imagine a different integration scheme to make the junction known as extension-first. On this case, the implantation is performed prior to the formation of the spacers. Thanks to the proximity to the oxide interfaces, we can perform a low-energy and high-dose implantation through the thin oxide liner that does not produces an amorphized region. Then the process follows the classical flow with the formation of spacers, source and drain epitaxy and SPER activation anneal.

The extension first method is attractive to improve the device performance for many reasons:

- The implantation step does not depend on the thickness of the source and drain epitaxies that presents a strong variability from wafer to wafer. Therefore, it is easier to define an optimal implant condition for every case.
- With low-energy implantation, the horizontal variability of dopant positions is reduced, thereby reducing the variations in electrical characteristics among different devices.
- A low energy implantation closer to the channel can improve the control of overlap position.
- Even in the case of overlapped devices, the LDD region can be more homogeneous with a sharp junction gradient.

2.4 IMPACT OF THE JUNCTION PROFILE ON DEVICE RELIABILITY.

On this session, the effects of overlap on the reliability of the low thermal budget devices are studied. BTI (Bias threshold instability) and HC (Hot Carrier) degradations of short channel devices ($L_G = 27nm$) fabricated within a low thermal budget are studied. Notably we address the impact of the junction position on the device degradation at high V_G and V_D stress conditions.

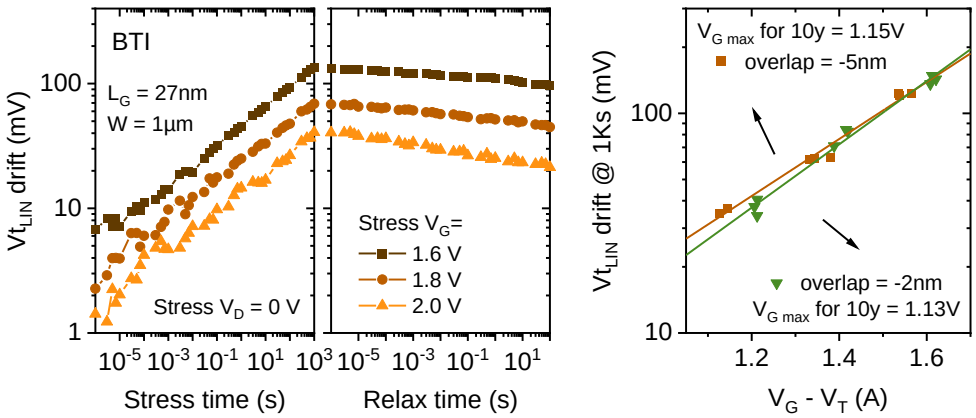


Figure 2-28 BTI degradation of 27nm devices with two different overlap conditions. (Left) Linear Vt shift during the stress and relax for 3 bias conditions. (Right) Degradation of linear Vt as the function of overdrive stress voltage for two different devices. TTF is evaluated for a 50mV V_T shift

Erreur ! Source du renvoi introuvable.-left illustrates the drift of linear Vt as function of time for a BTI stress. In this case, drain and source are grounded producing a vertical and uniform along the channel electrical stress field. The drift after 1ks is plotted against the overdrive stress bias for two similar LTB devices with different implantation conditions (**Erreur ! Source du renvoi introuvable.**-right). Both devices exhibit the same degradation at equivalent stress conditions, and similar maximum gate voltages for a 10 years lifetime. Even for short channel lengths, both variants are equally resilient to a vertical electrical field stress and the overlap position has no significant effect on the device reliability. Therefore we can conclude that the drift of Vt at $V_D = V_S$ is directly controlled by the quality of the gate stack.

In contrast, when the devices are tested under high drain current stress conditions, a different picture emerges. Figure 2-30 presents the drift of the saturation current over stress and relaxation time for variants S1 and S4 submitted to 3 $V_G = V_D$ stress conditions. At first glance, there is a clear difference between the HC and BTI degradation. For the former, we can evidence two distinct phases in the Id drift during the HC stress, whereas only one is visible for the BTI stress. In addition, each phase

appears dependent on the device overlap condition. The first degradation phase starts from the beginning of the test at 10µs up to 100 ms, and is more important for the S1 variant (low energy condition). The second degradation stage becomes noticeable after 100ms and is stronger for the S4 variant (strong energy condition). Similarly, the relaxation phases after stress are very different between HC and BTI tests. Figure 2-30-right presents the relative relaxation with respect to the last stress value. Interestingly, the relaxation kinetics are very similar during the first phase. This suggests that the same physical phenomenon is responsible for this recovery under BTI and HC stresses.

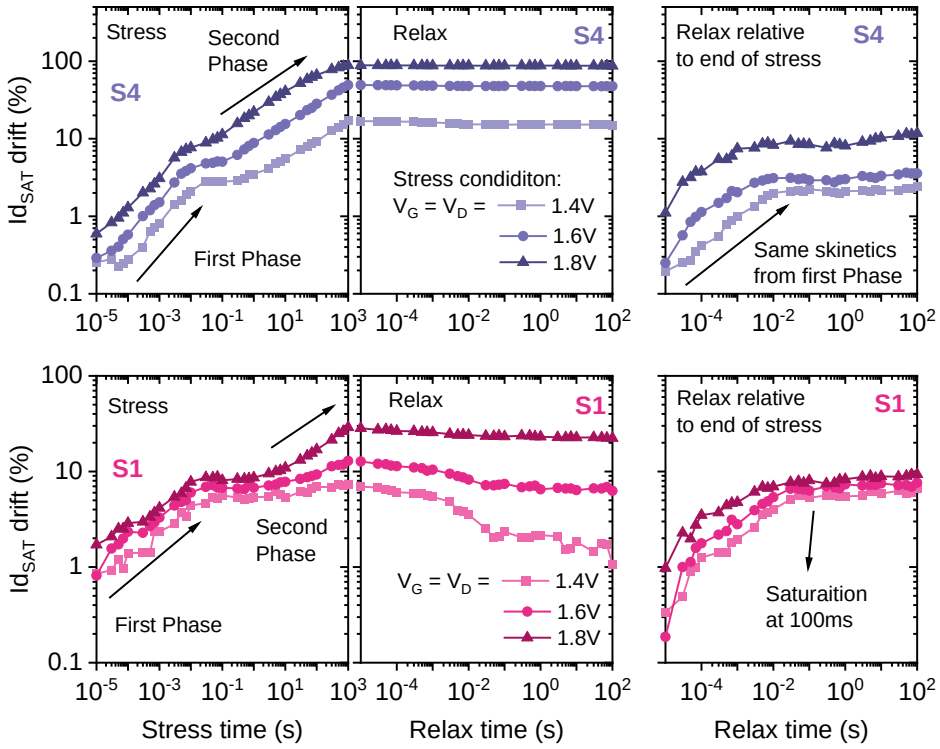


Figure 2-29 Comparison of HC degradation between variants S4 (top) and S1 (bottom). The stress is performed at same V_G and V_D on a 27nm gate length device. (Left saturation current degradation and revcovery for three bias conditions. (Right) Relative recovery in saturation current i.e. $Id_{sat}(t_{relax})/Id_{sat}(t_{stress}=1000s)-1$.

To determine precisely the mechanisms responsible for each phase, **Erreur ! Source du renvoi introuvable.** presents the total degradation for each variant as the function of the stress current at 10ks (**Erreur ! Source du renvoi introuvable.**-left) and 10ms (**Erreur ! Source du renvoi introuvable.**-middle) stress times. The difference in stress current level at same bias condition for each variant is consistent with the variation of access resistance with overlap. Although the degradation is proportional

to the current for each device, the complete dataset (all the variants) does not normalized over stress current. In addition, the dependence of the degradation with the implant conditions changes between 10ms and 1ks stress. For long stresses, the closer is the junction to the gate the higher is the degradation, whereas, for short stresses, underlapped devices exhibit a greater saturation current degradation. This result indicates that the stress current is not the single driver of the HC degradation in those devices. Actually, the second stage appears mainly driven by the stress current while the first one is related to the device overlap.

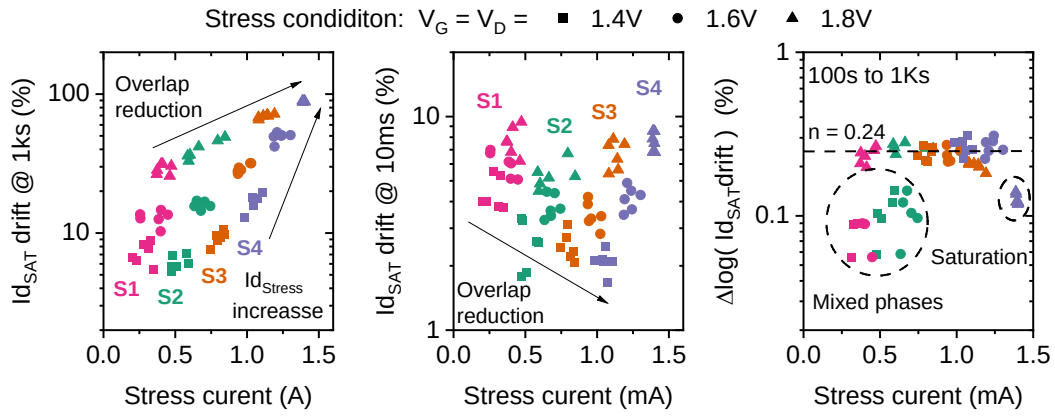


Figure 2-30 HC characterization of 27nm devices for S1-S4 variants versus the stress current. (Left) Degradation of the saturation current after a stress of 1ks. (Middle) Degradation of the saturation current after a stress of 10ms. (right) Time exponential factor (n) corresponding to the logarithmic difference of the saturation current drift from 100s to 1Ks .

From the previous results, it is therefore reasonable to consider two independent stress phenomena responsible for the complete HC degradation of LTB devices. To investigate this hypothesis a double power law model Eq. 2-13 is validated against the data. This model considers that each phase can be represented by a different saturated power law equation with different factors for the exponential dependence on bias and time. The total degradation is then given by the sum of the components.

$$\frac{\Delta I_{Dsat}(V_D)}{I_{Dsat0}} = \frac{1}{sat_1 \cdot {V_D}^{-n_{sat}} + A_1 \cdot {V_D}^{-\gamma 1} \cdot t^{-n1}} + \frac{1}{sat_2^{-1} + A_2 \cdot {V_D}^{-\gamma 2} \cdot t^{-n2}} \quad (2\text{-}133)$$

Fig x presents the degradation of the saturation current over time fitted for variants S1 and S4. On the left, the data is fitted with a classic power law model. The fit is clearly a bad representation of the devices degradation and is only an image of the drift from the second phase. On the middle, the data is fitted with the double power law model from eq. 2.13. The excellent match between data and model supports the hypothesis

of two independent stress phenomena. On the right, the data is subtracted from the first component of the fitted model of eq. 2.13. This results in a clear single power law dependence for the second phase. The time exponential factor (n) calculated at the end of the stress (**Erreur ! Source du renvoi introuvable.**-right) also confirms that the second phase is driven by a same mechanism regardless of the variants as a single value of n is obtained for all the variants when the second mechanism prevails over the first one.

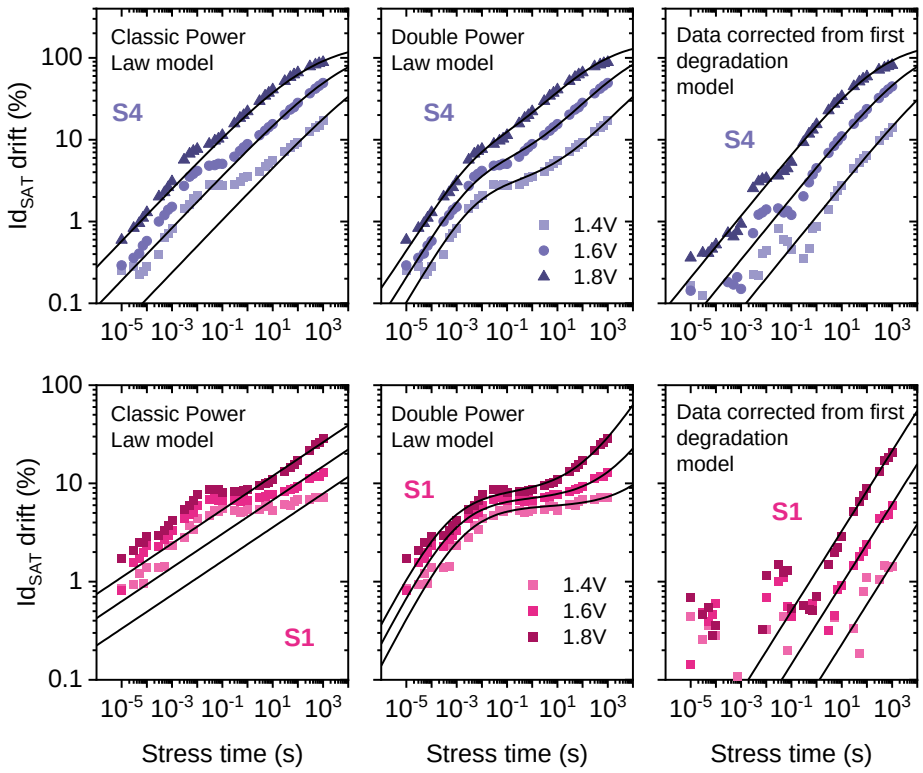


Figure 2-31 Comparison of power law models used to fit Id_{sat} degradation of variants S4 (top) and S1 (bottom) at 3 $V_G = V_D$ stress conditions. (Left) Classic and (Middle) double power law model fit for the Id_{sat} drift. (Right) Id_{sat} degradation subtracted of the first phase drift from the model.

From the fit of equation 2-13, it is then possible to evaluate accurately how the junction profile affects the HC reliability of the device. **Erreur ! Source du renvoi introuvable.**-left reports the weight of each mechanism in the degradation of the saturation current extrapolated at 0.04 years and $V_{DD} = 1V$ as the function of the overlap. As the junction moves closer to the channel, the weight of the fast component in the total HC degradation decreases while the long component increases. However, due to the early saturation of the first phase, the final degradation after a long stress period is mostly governed by the second mechanism. This effect is further visible in **Erreur ! Source du renvoi introuvable.** (middle and right). The Time To Failure at

$V_{DD} = 1V$ and the maximum V_{DD} for 0.04 year lifetime, both decrease with increasing overlap. This demonstrates that the first mechanism has finally a negligible impact on the final HC device reliability at nominal voltage: the fast mechanism saturates after a very short stress time and tends to relax when the stress is interrupted unlike the long term mechanism. Yet, note that, even if the fast mechanism has a little impact on long term device reliability, the short variations of drain current with time could affect the behaviour of the circuit if it is not taken into account during the design.

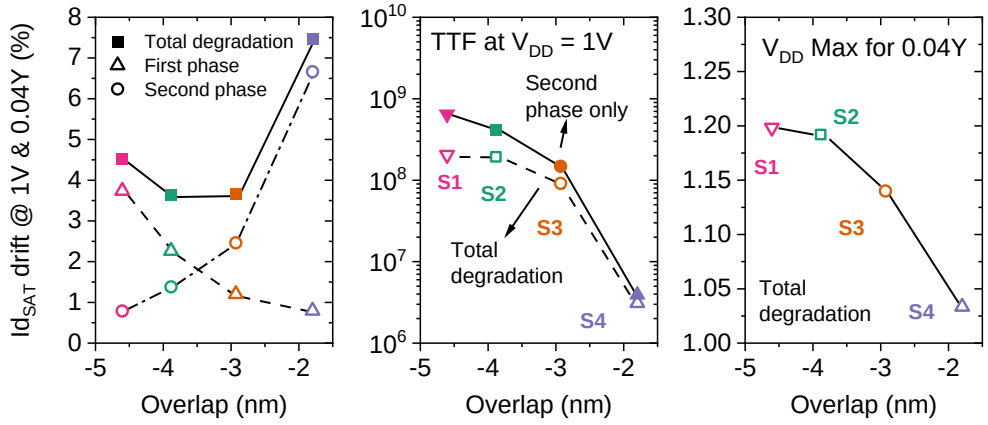


Figure 2-32 Comparison of the weight of fast (1) and long mechanism (2) on device reliability estimated from HC drift modeling for the variants S1-S4. (left) Extrapolated I_{dsat} degradation @V_{DD}=1V & 0.04y (Middle) Extrapolated Time to failure at $V_{DD} = 1V$ considering either the degradation due to the second phase or the total degradation. (Right) Maximum V_{DD} after 0.04 year of device working. The criterion for TTF and max V_{DD} is fixed to 10% I_{Dsat} degradation.

It is clear that the physical effect responsible the first phase is governed by the overlap position with respect to the spacer. As already seen in **Erreur ! Source du renvoi introuvable.**, the SiCO spacers present a strong trapping effect. It is also well known that the maximum generation of hot carriers in saturation regime occurs mainly in the vicinity of the channel/drain junction. Therefore, a possible scenario for the first degradation stage may be that, during the earliest instants of the stress, the spacer oxide captures a great amount of energetic carriers near the junction so that a depletion region is created under the spacers. This depletion region increases the access resistance and also shifts the threshold voltage of the device, thereby degrading the saturation current. After 100ms of stress times, the spacer is fully charged and the drift saturates. This hypothesis is supported by the recovery of the trapping observed during this phase, since the SiCO oxide is known to easily release its charges when the stress is interrupted. In addition, by overlapping the device, (1) the maximum of the HC generation is shift to the channel and (2) the electrical field due to the charged traps of

the spacers is screened out by the high doping density of the access. This, in fine, reduces the contribution of the first phase to the total current degradation.

The second phase itself is rather explained by a classic HC degradation of the device controlled by the drain stress current []. The closer the junction is to the channel, the smaller is the resistance of the source and drain. The reduction of access resistance increases the stress current at same bias, and thus the device degradation. In addition, this degradation during this second stage exhibits very little relaxation as depicted in **Erreur ! Source du renvoi introuvable.**. This well agrees with a standard HC degradation based on Si-H bond breaking [].

It is important to note that the first phase could also be induced by a degradation near the source. Indeed the electrical field between gate and source is similar to a BTI stress as the source is connected to the ground during the test. However, this hypothesis is not supported with the BTI results of **Erreur ! Source du renvoi introuvable.**. As concluded before, BTI degradation, due to the vertical electrical field from the gate, appears independent of the junction position.

Therefore we can summarize the physical phenomena governing each phase of the HC degradation on low thermal budget devices as follow:

- First phase: Fast capture and release of traps from the oxide material of the spacers. The trapping is induced by high energetic carriers near the drain junction on saturation and are released readily after the stress. The overlap shifts the position of hot carrier generation and screens out the effects of the spacer.
- Second phase: Device degradation is explained by a so-called MVHR (Si-H bond breaking) and is mainly driven by the stress current. Increasing the overlap reduces the access resistance, increases the stress current at same bias conditions, and in fine increase the HC degradation.

2.5 CONCLUSION

In this chapter, we proposed an improved model for the fringe capacitances of the FDSOI MOSFEFs. The model reproduces with great fidelity the simulation results from different overlap positions, channel length and gate oxide and silicon film thickness. Unlike previous models, this model can be applied to both underlapped and overlapped devices.

Using the fringe capacitance model, we proposed a novel non-destructive characterization method, named CV-AJP, to extract the junction profile from gate capacitance measurements. This characterization method allowed us to obtain the density of dopants and junction position on short length devices with an accuracy that cannot be reached by any other techniques like DIBL or SS. We validated the method against TCAD simulations on a 28nm traditional FDSOI MOSFET device and the extracted profile is able to perfectly reproduce the measured gate capacitance and subthreshold slope given by TCAD simulations.

The characterization method was applied to study the impact of drain and source implantation and activation on NMOS transistor performance. It allowed the identification of optimal implantation conditions for overlapped devices and the effects of low-temperature activation on access resistance, DIBL, and subthreshold slope. The method was also applied on PMOS transistors with a two-steps amorphization and doping implantation. The results showed that the activation of bore is not linked to the depth of the pre-amorphization step, and that a strong germanium implantation degrades the devices electrical performance. It was concluded that, although the germanium amorphization can improve the access resistance of the epitaxy region, it is not mandatory to activate the dopants near the junction.

From this study on the junction of both NMOS and PMOS transistors, it was also concluded that the thermal solubility of phosphorus and bore impurities is substantially greater near the oxides interface of the spacers and the BOX than the values obtained from the literature on bulk substrates. Therefore, the dopants can therefore be activated near this region without the need for an SPER recrystallization, which was so far a major drawback on the formation of the junctions of source and drain on LTB devices. From this result, an extension-first LDD without amorphization can be envisaged for improved device performance and lower variability.

Finally, the effects of the junction overlap on the Hot Carriers degradation of the device was studied. We discovered that two physical phenomena are responsible for the HC degradation at high V_{DD} stress: one fast and recoverable degradation due to the electron trapping in the oxide spacers, and a permanent degradation due to channel/gate oxide degradation. We showed that, the first contribution reduces with overlapped devices due to the screening of the spacer traps by the dopants while the second is mainly controlled to the level of the drain current and thereby increases with overlap.

Finally, record FoMs for LTB digital devices were presented for the first time on NMOS and PMOS devices. The results of this chapter were presented at the VLSI Symposium in 2022.

2.6 REFERENCES

Chapter 3: Characterization of low-K spacer Oxides

In most of MOSFET devices, the electric isolation between the gate and the source and drain is performed by the deposition of a thin dielectric layer after the gate patterning. Due to the proximity of the gate and the access of the device, the choice of this material has a big impact on the electrical properties of the transistors. To meet its isolation role, the material must present a low leakage current and a high breakdown voltage. On the other hand, it is preferred to have a low electrical permittivity material in order to reduce the parasitic gate capacitances and thereby improves the dynamic characteristics of the device.

Moreover, this oxide layer, named *spacer*, also plays different roles during the device fabrication. Firstly, the spacers are necessary to protect the gate stack from the following fabrication steps. In particular, the material must withstand the cycling process of low temperature source and drain epitaxy and wet cleaning [Jessy?]. The lack of this dielectric can lead to a lateral etching of the oxides of the gate stack itself during the epitaxy. Secondly, the material must encapsulate the Titanium Nitride (TiN) metal gate [Cao Minh Lu] to avoid its total consumption during the subsequent cleaning process steps. Finally, the dielectric thickness affects the depth of the ionic implantations in the source and drain region and the final junction position as discussed in chapter 2: as reminder, on a low temperature integration process, due to the negligible dopant diffusion at 500°C, the junction position is directly given by the position of dopants after implantation. Therefore, the oxide must be sufficiently robust to etching in order to minimize variations of its thickness before the implantation steps and the resulting variability of the overlap position on small devices across the whole wafer.

The choice of this dielectric material is therefore critical. It must own all the right aforementioned electrical and chemical properties, in addition to be compatible with a low thermal budget T<500°C fabrication process. After a full benchmark of different low-k dielectrics materials (SiN, SiOCN, SiCO and SiCBN) by [Cao Minh Lu] in term of mechanical, chemical and electrical properties, the SiCO oxide was selected as the

best candidate for the spacer oxides for a low temperature process. Its main advantages are (1) a high etching robustness whether the chemistry used for the wet cleaning process and a great selectivity to anisotropic etching (2) a great deposition uniformity on patterned gates and (3) a low permittivity of 4.5, which is much smaller than the permittivity of the traditional spacer material SiN ($\varepsilon r_{SiN} \approx 7$).

On this chapter, we will perform a further evaluation of this material with a greater focus on its trapping characteristics. To achieve a great characterization of this electrical effect, a new and original advanced measurement method has been developed. Therefore, this chapter is divided in 3 parts. Firstly, we will present the new ultra-fast capacitance measurement method. Then, we will investigate the different trapping mechanisms in this material and how they can be tuned by hydrogen or nitrogen anneals. Finally, we will explore how the dielectric gate spacer can affect the electrical FoMs of the device.

3.1 TRAP CHARACTERIZATION

3.1.1 Interface traps

Interface traps are crystal defects localized at the interface between the Silicon substrate and the interfacial oxide. The lattice mismatch between the crystalline structure of the silicon substrate and the amorphous oxide induces intrinsic strain at the interface responsible for the defect generation. The most common defects are Pb centers that can be easily revealed through Electron Spin Resonance measurements [A Stesmans]. Different types of Pb centers were identified with different morphologies, but the most common ones are the Pb0 and Pb1 centers, as illustrated in **Erreur ! Source du renvoi introuvable.**.

Those defects create energy states within the silicon bandgap and are often considered as amphoteric. This means that each microscopic defect induces two energy levels in the band gap: a donor state (+/0) in the lower part of the bandgap and an acceptor state (0/-) in the upper part. A single defect can then capture a hole or an electron depending on the position of the Fermi level at the semiconductor interface. More precisely, the filling of the states created by interface traps is given by the integral of the Fermi function multiplied by the trap density of states. In a simplified manner, we approximate the Fermi function by a Heaviside function and then consider

that a trap captures an electron when the Fermi level lies above its energy state and lost this electron when it lies below.

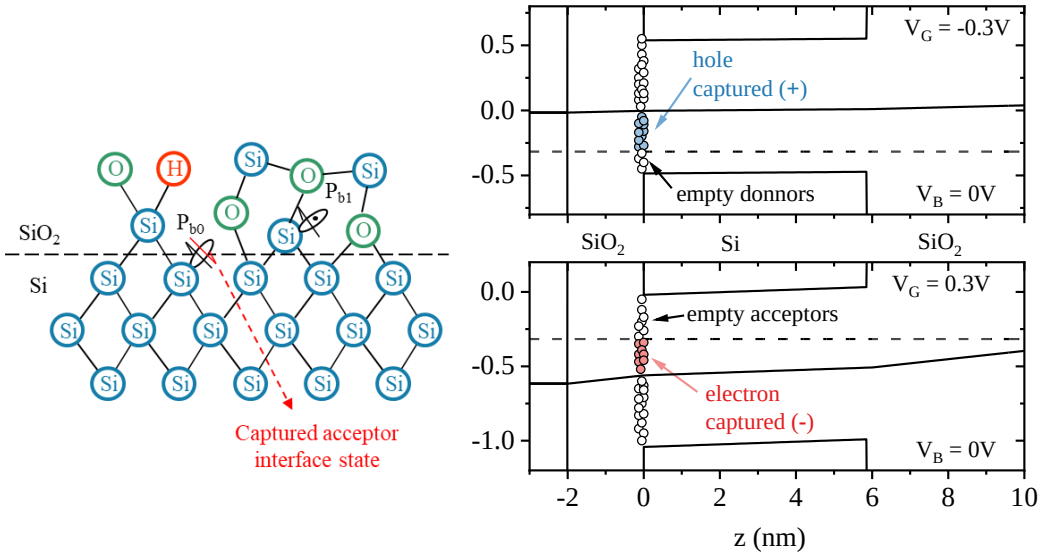


Figure 3-1 Sketch illustrating the concept of interface states. (Left) Pb centers are silicon band-gap defects that originate from the lattice mismatch between the oxide and the substrate. When the Fermi level lies above the energy of an acceptor trap, it can capture an electron at the interface. (Right) Sketch illustrating the filling of the Pb centers on a FD-SOI structure. Acceptor traps capture electrons and can be at the charged states (0/ -1). While donor traps capture a hole and can be at the states (0/+1).

In the case of a Pb center, when the Fermi level lies in the upper part of the bandgap, its acceptor state can be either filled (-1) or empty (neutral =0) whereas its donor state remains always filled by an electron i.e neutral (0) as depicted Figure 3-1. On the other hand, when the Fermi level lies in the lower part of the bandgap, its acceptor state remains always empty (neutral = 0) whereas its donor state can be positive charged (+1) when it lies above the Fermi level and neutral (0) when it lies below. Note that the framework proposed here for Pb centers can be easily generalized to other possible microscopic interface defects e.g N-induced interface defects [GARROS, CASSE] that are not necessarily amphoteric. In that case, we simply consider a single energy state per trap either donor or acceptor.

The dynamic of filling and release of interface traps is often described by the SRH (Shockley Read Hall) model [ref]. It models the probability of capture and release of traps in the bandgap as the function of the Fermi level and the energy of the trap. For acceptor, it considers that the capture probability of a given trap is proportional to the product of the number of un-occupied interface states and of the density electrons at the Si interface, while its release or emission probability is only proportional to the

number of occupied interface states. On thermal equilibrium, the total number of captured traps is given by the state where the probability of capture and emission of each trap is the same.

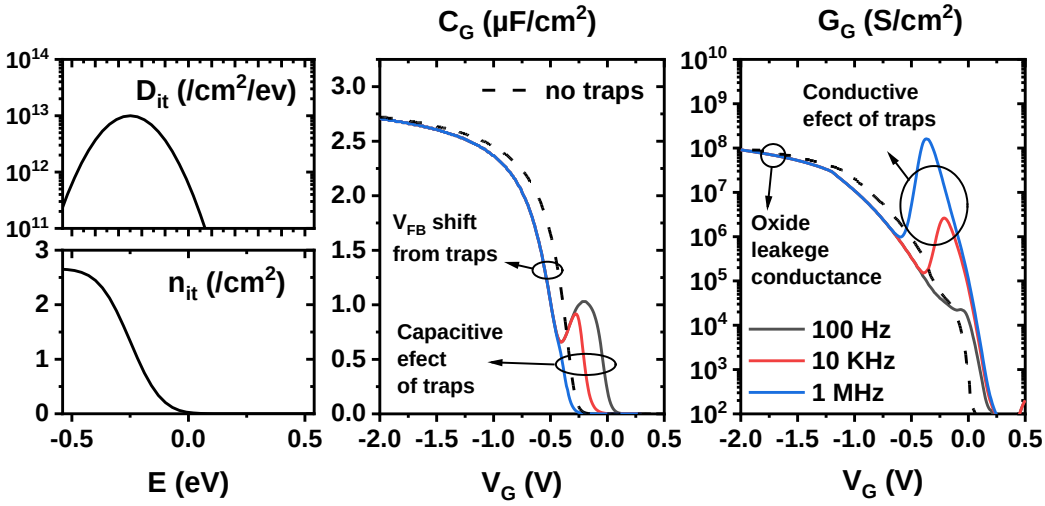


Figure 3-2 Typical Capacitive (C_G) and Conductance (G_G) versus V_G characteristics of a p-doped substrate MOS capacitor with a 1nm oxide containing a high amount of interface traps. The curves at at various frequencies are calculated using a Poisson-Schrodinger CV simulator. (left top) Trap profile D_{it} over the bandgap. (left bottom) Total trapped charge nit as function of the Fermi level. (center) Effect of those traps on the CV characteristic at different frequencies. (right) Effect of traps on GV at different frequencies. Besides the shift of V_{FB}, the interface traps induces a typical capacitive and conductive response that is dependent on frequency.

The trapping phenomenon affects the electric characteristics of the MOS device in a couple of ways as illustrated in **Erreur ! Source du renvoi introuvable.**. The charging and discharging of those interface traps as the function of the surface potential generates a capacitive and conductive frequency dependent response to the A/C signal applied to the gate bias. In addition, the net positive charge contained in those traps modifies the relationship between the surface potential and the gate voltage $\psi s(V_G)$ as despite in equation (X), that induces a stretch of the CV characteristics along the V_G axis as well a shift on the flat band voltage V_{fb}. The capacitive effect of the traps can be modelled as a capacitance component (C_{it}) in parallel with the semiconductor (C_s) that are both in series with the oxide capacitance (C_{ox}) as depicted in equation (X). In the case of an interface state continuum and in absence of potential fluctuations the capacitive effect of the traps are given by eq. (X). In addition to the capacitive effect, the trapping and emission of traps induced by the AC signal from the capacimeter induces losses that are modelled as a conductive component of those traps (G_{it}) as described by eq. (X).

$$V_{G0} = \Psi_{S0} + V_{FB} - \frac{Q_{SC}(\Psi_{S0}) + Q_{trap}(\Psi_{S0})}{C_{ox}} \quad (X)$$

$$\frac{1}{C_{tot}} = \frac{1}{C_{ox}} + \frac{1}{C_{it} + C_s} \quad (3\text{-}1)$$

$$C_{it} = q \cdot N_{it} \frac{arctan(\omega \cdot \tau_m)}{\omega \cdot \tau_m} \quad (3\text{-}2)$$

$$\frac{G_{it}}{\omega} = \frac{q \cdot N_{it}}{2 \cdot \omega \cdot \tau_m} ln(1 + (\omega \cdot \tau_m)^2) \quad (3\text{-}3)$$

The extraction of the interface trap density on MOS capacitor then requires the measurement of the device admittance at various frequencies and biases $Y(V_G, \omega)$ using a standard impedance meter e.g, Keysight 4291. Several well-known characterization based on CV or GV signals are used to extract the Dit profile. The three most famous are the Conductance method and High-Low frequency method

The conductance method [Nicollian] uses the GV peaks measured in the depletion regime at each frequency in order to extract the Dit profile and the characteristic capture time of each trap over gate bias. It is very sensitive method able to detect trap density down to 10^9 defects per cm^2. However, it can only be used to probe traps that "resonate" at the frequencies supported by the capacimeter. This exclude very fast traps near the band edges and very slow traps that are located deeper into the oxide. To extract the density of traps using this method, the conductance G_{it} associated with the interface traps that are active at a given V_G is calculated from the total measured capacitance and conductance as function of the frequency as described on eq. X. Then the interface trap density N_{it} can be determined from the magnitude of the maximum of G_p/ω while the time constants τ_m can be determined from the frequency position (f_{max}) of the maximum of G_p/ω curve.

$$\frac{G_{it}}{\omega} = \frac{{C_{ox}}^2 \cdot \omega \cdot G_{mes}}{{G_{mes}}^2 + (C_{ox} - C_{mes})^2 \omega^2} \quad (3\text{-}4)$$

$$N_{it} = \frac{4}{q \cdot ln(5)} \left(\frac{G_p}{\omega}\right)_{max} \quad (3\text{-}5)$$

$$\tau_m \approx \frac{1}{\pi \cdot f_{max}(V_g)} \quad (3\text{-}6)$$

The High-Low frequency or Terman's method uses two capacitance measurements at a High and a Low frequency to extract the Dit profile. The technique considers that at low frequency, all the interface traps can follow the gate AC small signal, producing a maximum Dit response in the CV curve. On the other hand, at high frequency, the signal is too fast for the traps, producing a null C_{it} capacitance. The method is simple but suffers from the same limitations than the conductance method. For both methods, the positioning of the traps in the gap can be extracted from the band-bending that can be calculated using the Berglund integral. It allows obtaining the surface potential as function of the gate bias where each trap is extracted. $\emptyset_s^0 = 0$ when V_g^0 is referenced at flat band voltage.

$$D_{it}(V_g) = \frac{C_{ox}}{q}\left(\frac{C_m^{lf}}{C_{ox} - C_m^{lf}} - \frac{C_m^{hf}}{C_{ox} - C_m^{hf}}\right) \tag{3-7}$$

$$\emptyset_s = \emptyset_s^0 + \int_{V_g^0}^{V_g}\left(1 - \frac{C_m^{hf}(V)}{C_{ox}}\right)dV \tag{3-8}$$

Another technique to probe the interface trap density is based on Terman's method [ref]. This technique uses the stretch of a high frequency (~1MHz) CV curve along the V_G axis to derive a trap profile over Si bandgap. The density of traps are extracted as function of the stretch of the surface potential using eq. (X). It compares the band-bending induced by traps with the variation of surface potential with the gate bias from an ideal case without traps. The advantage of this technique is that it only use the stretch of CV curve and do not rely on the resonance of the traps. Therefore, it can extract traps that are too slow to be measured with the previous methods. Nonetheless, it requires a reference capacitance curve that represents an ideal device without any traps. This reference curve can be obtained from an ideal devices or from the simulation of the device that takes in account its physical parameters.

$$D_{it}(\emptyset_s) = \frac{C_{ox}}{q}\left[\left(\frac{d\emptyset_s}{dV_g}\right)^{-1} - 1\right] - C_s(\emptyset_s) \tag{3-9}$$

Each extraction method from direct measurement has its advantages and limitations. Therefore, for a reliable extraction of the oxide interface features, in the following study, the measured capacitance and conductance curves will also be fitted against a 1D Poisson-Schrodinger CV simulator that models the trap response.

3.1.2 Oxide traps

Oxide traps are defects that are localized inside the oxide volume and are mostly independent on the substrate material at the interface. This time, they generate energy levels within the oxide bandgap that can be filled by an electron or hole (**Erreur ! Source du renvoi introuvable.**). However, unlike interface traps, the filling probability of those defects will also depend on their depth within the oxide, based on the fact that the trap capture occurs through a tunnelling mechanism between a carrier of the Si interface and the trap energy level.

As a result, the charging of those traps happens in a much slower rate than interface trap, inducing no AC effects on classic capacitance measurements. The charge variation impacts only on the V_{fb} shift and depends on the defect position within the oxide. They can be properly characterized using noise measurements on very small MOSFET devices or by TDDS method [Grasser]. However, due to the necessity of specific devices, this technique remains out of the scope of this work. Each one of the analysis of the LTB spacer oxide material is performed here in simple MOS capacitors.

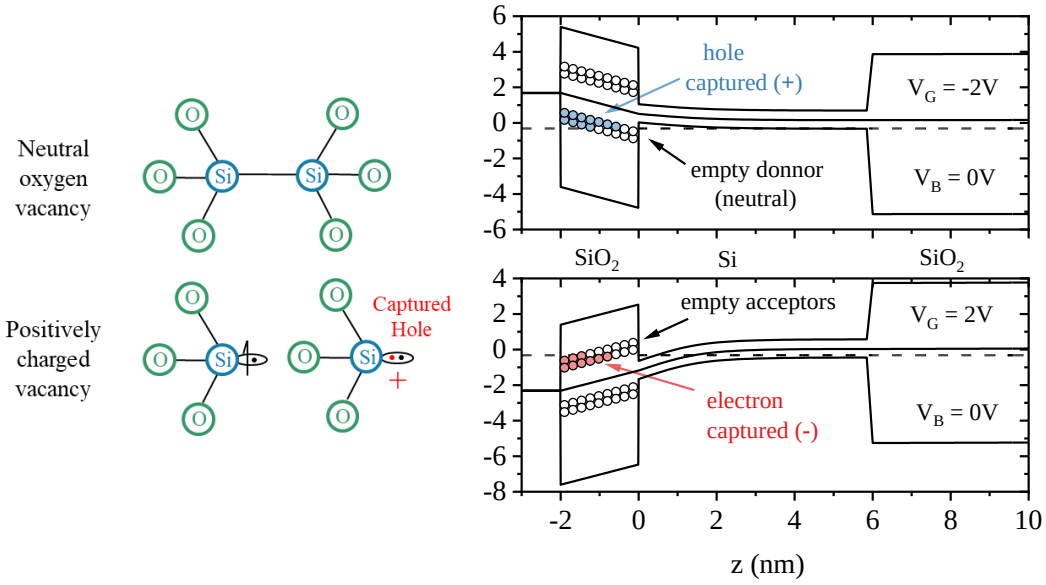


Figure 3-3 Sketch illustrating the concept of oxide traps with the example of oxygen vacancies. (Left) oxygen vacancies, oxide trap on a neutral and passively charged state after capturing a hole. (Right) Sketch illustrating the filling of oxide traps on a FD-SOI structure. Similar to interface states, oxide traps can be either donors or acceptor. The main difference been that filling probability of those defects will also depend on their depth within the oxide.

On MOScaps, this trapping can be observed by performing simple CV hysteresis measurements. It consists in measuring a double sweep CV curve: a sweep from inversion to accumulation followed by a sweep from accumulation to inversion. The

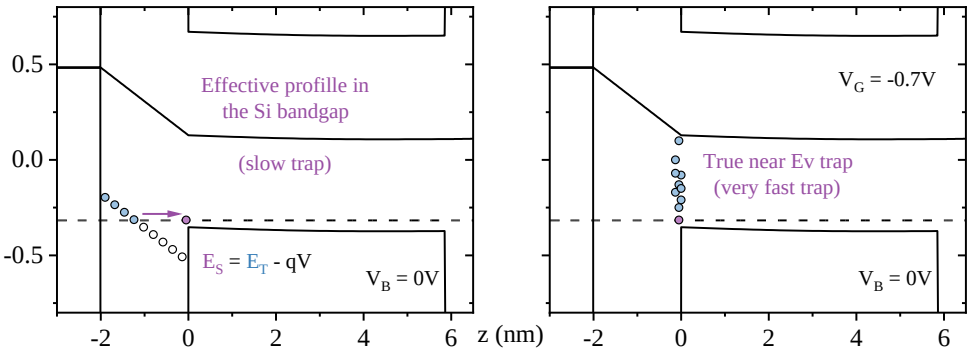
oxide trapped charge is then given by the difference in V_{fb}between both curves. However, this method is far from reliable as the extracted trapped charge is highly dependent on the time to perform the measurement.

A more advanced technique used to investigate oxide traps is based on a Measure-Stress-Measure (MSM) method. For MOScaps, this involves measuring the shift of the flat band as a function of the time the device is stressed at a specific gate bias. The MSM method follows these steps: (1) A first capacitance measurement is performed at t_0. (2) The device is stressed by applying a strong gate bias that allows the oxide traps to traps the carriers from the semiconductor. (3) Another capacitance measurement is performed at t_1 to probe the shift in V_{fb} induced by the captured charges. (4) Steps 2 and 3 are then repeated with increasing time lengths until t_n. The major limitation of the MSM method is that, in order to determine the shift of V_{fb} between stresses, the measurement needs to be performed at low values of V_G. However, due to the long time required to obtain the CV curve by most capacimeters, a significant percentage of the traps are released during each measurement. This can potentially invalidate or hide the contribution of traps with a time constant that is smaller than the time required to obtain the full CV curve. Nonetheless, the stress can also induce the formation of interface states that can also be responsible for a shift on the flat-band voltage.

3.1.3 Slow oxide border traps vs fast Near-Band Si traps

Even though oxide border traps and interface traps have different physical characteristics and origins, some characterization methods are not limited to probing only one type of trap. The most extreme case of this phenomena happens in the differentiation between border traps and Near-Band Si traps. Border traps are oxide traps that are near the oxide semiconductor interface, they are too slow to induce a capacitive signal when measured even at low frequencies but are still fast enough to be captured and induce a shift of V_{fb} when obtaining the CV curve. Near-band Si traps, on the order hand, are interface silicon traps that are located close to the conduction or valence bands. As a result, their response times are very fast, and they are unable to generate a conductive signal during measurement. In addition, they produce a capacitive signal even at high frequencies, making it impossible to probe these traps using traditional H-L frequency and conductance methods. For those traps, the

Therman's method based on the stretch of the flat-band curve is required. However, both traps can present the same signature when extracted with this method. As depicted at Figure X, both traps are captured at the same surface potential, because the stretch of the flat-band do not provide information about the time behaviour of the trap. It is important to account for this oxide trapping during the characterization of the interface traps itself. Inaccuracies in Dit extraction can arise from the fact that the total V_{FB} shift is the sum of trapped charges at the interface and inside the oxide. In the next section, we will introduce a new fast CV method that enables proper differentiation between these two types of traps.



Sketch illustrating the difference between oxide border traps and Near-Band Si traps and how they can provide the same signature on the gap. (Left) Slow Oxide traps that are near the border can be captured from inside the gap and present an effective profile similar to interface traps. (Right) Very fast near silicon traps near valance band traps.

3.2 ULTRA FAST CV METHODS

As previously discussed, CV characteristics of MOS capacitors (MOScap) are classically measured in few seconds using a capacimeter. Yet this value of few seconds limits the range of useful characterization that can be performed on these devices. For instance, the conventional MSM BTI reliability characterization is limited by the unavoidable recovery during every measurement step [refs]. Likewise, it is almost impossible to characterize properly the hysteresis effect present in some oxides when the time to measure the capacitance is much larger the typical time constant of the phenomenon. Furthermore, due to a strong hysteresis effect that will be discuss later, the characterization of the oxide material used for the spacers of LTB devices requires a different method.

That is why, ultra-fast (< 10 µs) CV ramp were recently explored on MOSFETs by [refs] but not on pure capacitor, for which they are more suitable. Therefore, we

propose in this thesis new ultra-fast and powerful CV techniques for MOScap, able to overcome the time limitations of former techniques. The first part describes how the technique is relevant for an interface trap spectroscopy. Then, an ultra-fast CV-based BTI technique is highlighted.

3.2.1 Ultra-Fast CV measurement patterns

Two methods are explored for fast capacitance (< 10µs) characterization: CV_{pulse} and CV_{ramp}. Both are performed using a Keysight B1530 Wave Generator Fast Measurement Unit (WGFMU) in order to simultaneously apply precise signal patterns and perform current measurement with a 10ns sample interval resolution. The patterns are applied to the gate of a MOScap and the substrate is connected to the ground. The principle of both capacitance measures is illustrated in **Erreur ! Source du renvoi introuvable.**-left.

The CV_{pulse} method is performed using several small and sharp (100ns) pulses that are applied on the device gate (**Erreur ! Source du renvoi introuvable.**-right-top). The pulses have a ΔV_G amplitude around a V_{G0} bias and are applied on both directions to separate trap charging and discharging phases. Then, the current is sampled from the beginning of the pulse to a period long enough to allow the device being completely stabilized i.e. $I_G(t) \sim 0$ (see fig.2 top). The pulse pattern is then repeated for different V_{G0} biases to extract the $C_{pulse}(V_{G0})$ derived from the integral of $I_G(t)$ over time as presented on eq. 3-1.

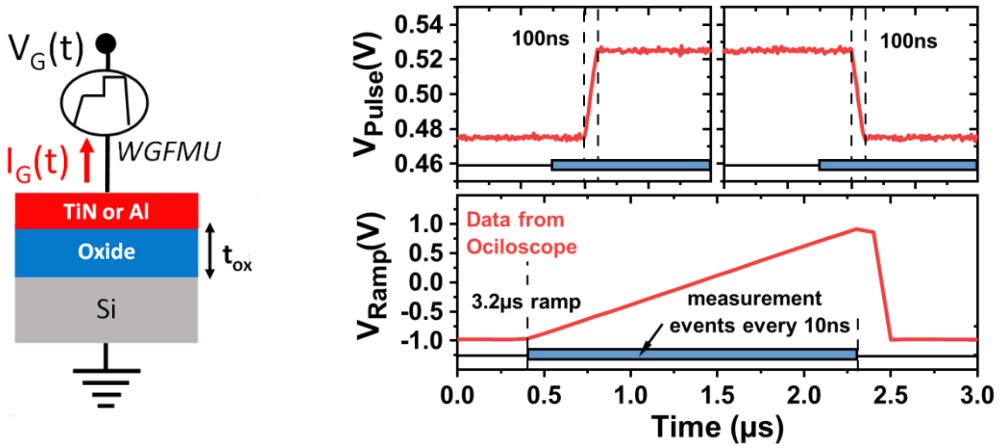


Figure 3-4 Oscilloscope traces of the two signals used for the fast capacitance methods (Top) Two short 100ns pulses of 50mV (Bottom) 3.2µs ramp. The current measurement events are illustrated by the blue bar.

$$C_{pulse}(V_{G0}) = \int_{t0}^{t} \frac{I_G(t)}{\Delta V_G} dt \tag{3-10}$$

$$C_{ramp}(V_G) = I_G \left(\frac{dV_G}{dt}\right)^{-1} \tag{3-11}$$

CV_{ramp}, on the other hand, is performed using a longer V_G ramp (1 to 5µs/V) covering all bias of interest (**Erreur ! Source du renvoi introuvable.**-right-bottom). The capacitance is then derived from displacement current I_G measured during the ramp using eq. 3-2. For both methods, the parasitic capacitance of the WGFMU measurement unit is subtracted from the raw data. This is done by performing an OPEN measure, without any electric contact to the device. In this OPEN measurement, the probes are close but not in contact to the DUT where a pulse or ramp measurement is performed. The capacitance obtained from this measurement is later subtract from the measurements of the DUT capacitance to remove the parasitic capacitance.

The curent response of the pulsed and ramp pattern are given as by the following equations. The capacitance is considered constant on both cases and the gate conductance is sufficiently small to not affect the signal. The variable C_{tot} is the sum of parasitic capacitance components from the WFGMU, the cables and the probes with the capacitance from the DUT. R_S is the series internal resistance from the WFGMU curent prober. On the ramp signal, ΔV_r represents the signal sweep voltage and ΔT its duration. It is clear that the signal has a transitory phase given by the exponential component with a time constant of τ_{RC}. This component becomes null when $t \gg \tau_{RC}$ the curent becomes constant an proportional to the total capacitance. As a rule of thumb, when $t > 5\tau_{RC}$, the value of capacitance can be calculated using eq X.

Similarly, the pulse signal, is composed of a first fast ramp signal of magnitude ΔV_p and period t_{rise} followed by a constant stabilization phase, the charge is calculated from the integral of the curent over time as given by eq. X. On this case, the signal also has an exponential amortised transitory phase before the curent zero and the charge becomes constant.

$$I_{ramp}(t) = \frac{\Delta V_r}{\Delta T} \cdot C_{tot} \cdot \left(1 - e^{\frac{-t}{\tau_{RC}}}\right), \qquad for\ x > 0 \tag{3-12}$$

$$\lim_{t\to\infty} Q_{pulse}(t) = \frac{\Delta V_G}{t_{rise}} \cdot C_{tot} \tag{3-13}$$

$$I_{pulse}(t) = \begin{cases} \dfrac{\Delta V_p}{t_{rise}} \cdot C_{tot}\left(1 - e^{\frac{-t}{\tau_{RC}}}\right), & x < t_{rise} \\ \dfrac{\Delta V_p}{t_{rise}} \cdot C_{tot}\left(e^{\frac{t_{rise}}{\tau_{RC}}} - 1\right)e^{\frac{-t}{\tau_{RC}}}, & x \geq t_{rise} \end{cases} \quad (3\text{-}14)$$

$$Q_{pulse}(t) = \begin{cases} \dfrac{\Delta V_p}{t_{rise}} \cdot C_{tot}\left[\tau_{RC}\left(e^{\frac{-t}{\tau_{RC}}}\right) + t\right], & x < t_{rise} \\ \dfrac{\Delta V_p}{t_{rise}} \cdot C_{tot}\left[t_{rise} + \tau_{RC}\left(e^{\frac{t_{rise}}{\tau_{RC}}} - 1\right)e^{\frac{-t}{\tau_{RC}}}\right], & x \geq t_{rise} \end{cases} \quad (3\text{-}15)$$

$$\tau_{RC} = R_S C_{tot} \quad (3\text{-}16)$$

$$\lim_{t\to\infty} Q_{pulse}(t) = \frac{\Delta V_G}{t_{rise}} \cdot C_{tot} \quad (3\text{-}17)$$

Erreur ! Source du renvoi introuvable. illustrates the measured displacement gate current (I_G) from a 100ns CV_{pulse}, applied to a 3nm HfO_2 oxide annealed by Forming Gas Anneal (FGA), that contains a very small amount of interface states (<10^{10}/cm^2). The total charge $Q_G(t)$ is calculated from the integral of the measured current $I_G(t)$. Albeit the applied pulse has a short duration, the internal RC circuit of the WGFMU limits the charge stabilization time. The capacitance is therefore evaluated after $2\tau_{RC} \sim 300ns$.

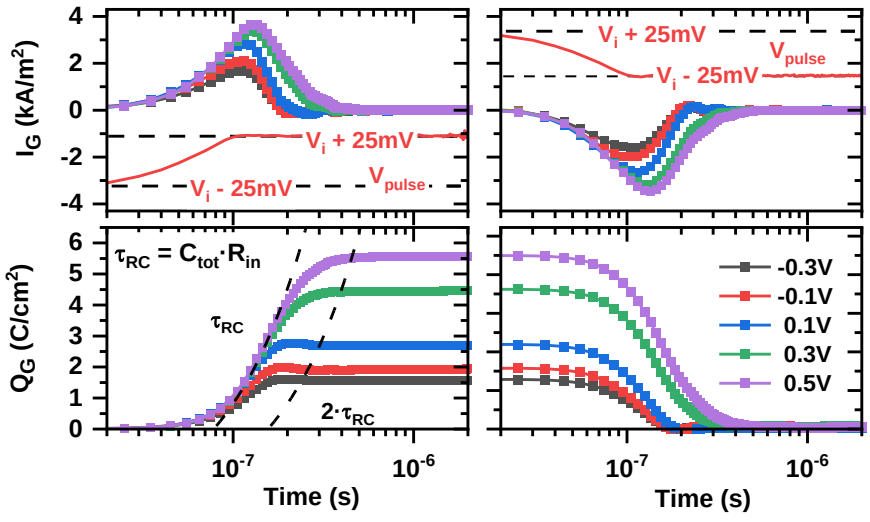


Figure 3-5 Results of CV_{pulse} on a 3nm HfO_2 oxide w/o D_{it}. (Top) Measured current (I_G) over time for 50mV pulses around V_G. (Bottom) total charge Q_G. The capacitance is evaluated after the charge stabilization (> $2\tau_{RC}$).

Erreur ! Source du renvoi introuvable.-left and middle shows that the CVs measured with the two techniques are finally identical to the one obtained using a classic capacimeter. Yet, the benefit of both methods is obvious as the complete CV can be obtained under 10µs, instead of a few seconds for the capacimeter. A last point that must be emphasized is that CV_{ramp} characteristics can be erroneous if the voltage ramp is too fast (**Erreur ! Source du renvoi introuvable.**-right), due to the intrinsic delay τ_{RC} formerly discussed. This delay never appears for CV_{pulse} since the measure is made after signal stabilization. That is why this last technique will be preferred for trap spectroscopy.

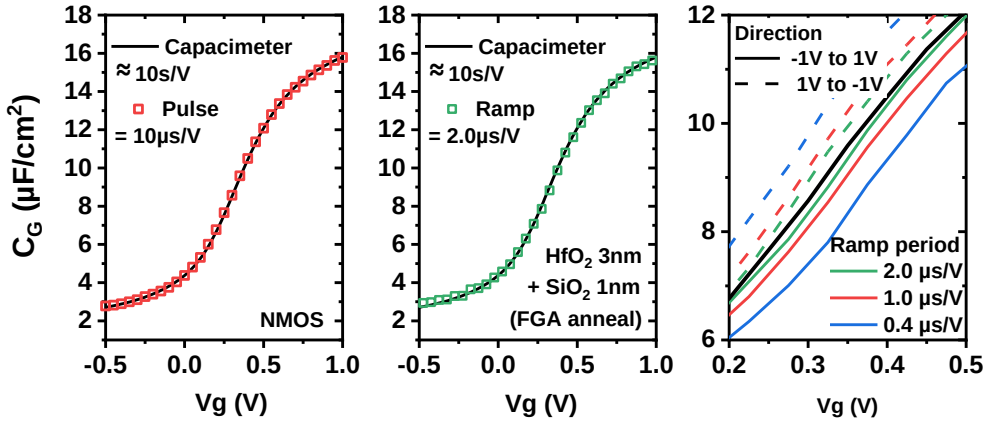


Figure 3-6(Left) Comparison of CV characteristics of a 3nm HfO2 oxide with low density of D_{it} using CV_{ramp}, CV_{pulse} and a classic capacimeter (Right) CV_{ramp} variability vs ramp speed and directions. Ramp speed must be greater than 2µS/V to recover the right CV curve.

3.2.2 Interface States Spectroscopy using pulsed pattern

Due to the fast 10ns sample interval, the CV_{pulse} technique allows to obtain not only the capacitance characteristics at a given V_{G0} but it also allows to separate the response of the semiconductor free carriers from oxide interface state charges. Interface states have a different capture and release times that are not correlated to the MOScap τ_{RC}. Therefore, the $Q_G(t)$ characteristics of oxides with high amounts of D_{it} can contain two charging periods, a first limited by the measurement setup and a second related to the exchange of charges with interface & oxide defects (**Erreur ! Source du renvoi introuvable.**).

This technique can only be applied on accumulation regime, as the inversion of carriers is generally not visible in simple MOS capacitors. Capture and emission pulses

are dependent on the type of interface states that are been characterized: For donor traps (+/0), capture pulses are applied from positive to negative values while emission pulses are applied from negative to positive values of V_G. For acceptor traps (0/-), the capture and emission periods pulses are applied on the other direction.

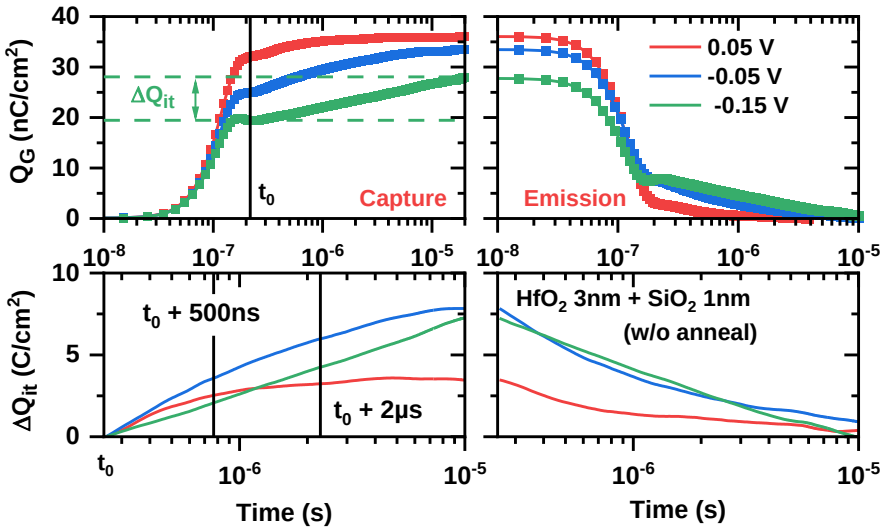


Figure 3-7 CV_{pulse} technique applied on a 3nm HfO2 oxide w/o FGA anneal. (Top) Extraction of $\Delta Q_{it}(t)$ from $Q_G(t)$ due to interface traps during capture and emission steps at various V_{G0}. (Bottom) ΔQit dynamics after the initial charging phase due to displacement current ($t > t_0$).

As, for a given V_{G0}, the Fermi level lies on a specific trap energy level in the Si bandgap with an uncertainty of kT/q i.e. $E_T \pm kT/q$, it is possible to use this specific measure to make an in-depth D_{it} spectroscopy. More precisely, it is possible to extract (1) the D_{it} profile over bandgap derived from the interface trap charge by derivation i.e. eq. 3-3 and (2) the characteristic trap capture $\tau_C(E_T)$ and emission time $\tau_E(E_T)$, each one extracted independently (Fig. 4 (bottom)).

$$D_{it}(t, V_{G0}) = \frac{C_{it}(t, V_{G0})}{q} = \frac{\Delta Q_{it}(t)}{q \Delta V_S} \tag{3-18}$$

Erreur ! Source du renvoi introuvable.-left compares the extracted D_{it} profile at several charging times to the one obtained from the conductance method $G(\omega)$ [6]. Both methods are perfectly consistent and effective to extract the D_{it} profile close to midgap $[E_i, E_i + 0.2\ eV]$. Moreover, CV_{pulse} also allows an effective reconstruction of the CV characteristic at different capture or release times (**Erreur ! Source du renvoi introuvable.**-right). However, above $E_i + 0.2eV$, the CV_{pulse} method

underestimates the true D_{it} density that can be extracted from an accurate modelling of the CV, which accounts for both the C_{it} & CV high frequency stretch-out [7, 8].

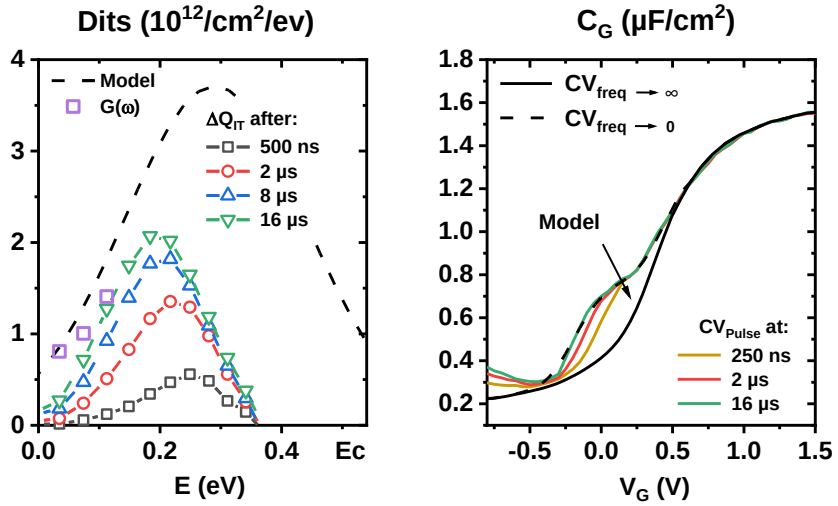


Figure 3-8(Left) D_{it} profile extracted from ΔQ_{it} at different capture times t-t0 vs. values obtained from G(ω) method, and, from an accurate modeling of the C(ω) vs $G(\omega)$ network. (Right) CV characteristic extracted from CV_{pulse} method at different capture times.

This is due to the fact that a significant amount of these traps, filled before the $2\tau_{RC}$, are not considered in $\Delta Q_{it}(t)$. Next, to extract τ_C and τ_E, we perform a temporal simulation of $Q_G(t)$ under various capture and emission pulses (**Erreur ! Source du renvoi introuvable.**). To do it, we combine the D_{it} profile formerly extracted (dotted line) to a trapping/de-trapping simulator similar to the one presented in [9]. The continuous $D_{it}(E)$ profile is meshed in a discrete number of traps $N_{it}(\mathrm{E_i})$ over the bandgap. For a given $V_G(t)$ signal, the simulator calculates the variation of the filling rate (f_R) of each trap i at each time t by solving equation 3-4. where $\tau_{c,i}(t)$ and $\tau_{e,i}(t)$ denote the capture and emission time constants derived from Shockley–Read–Hall (SRH) model. $\tau_{c,i}$ and $\tau_{e,i}$ $\Delta V_{FB}(t)$ and $Q_G(t)$ are then computed at each time t according to equation 3-5.

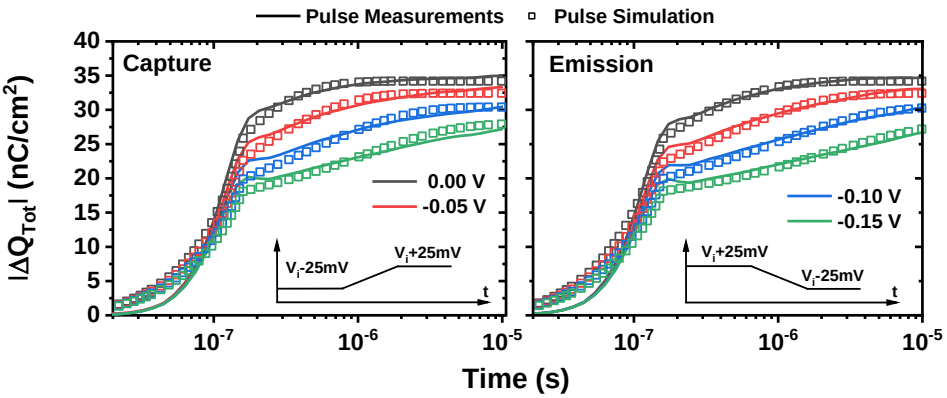


Figure 3-9 Simulation of $Q_G(t)$ transients from CV_{pulse} using a trapping/detrapping "SRH-like" model and the D_{it} profile obtained from Fig. 5. The model well captures the whole trapping dynamics with a unique set of constant times (τ_e & τ_c.) at each bias.

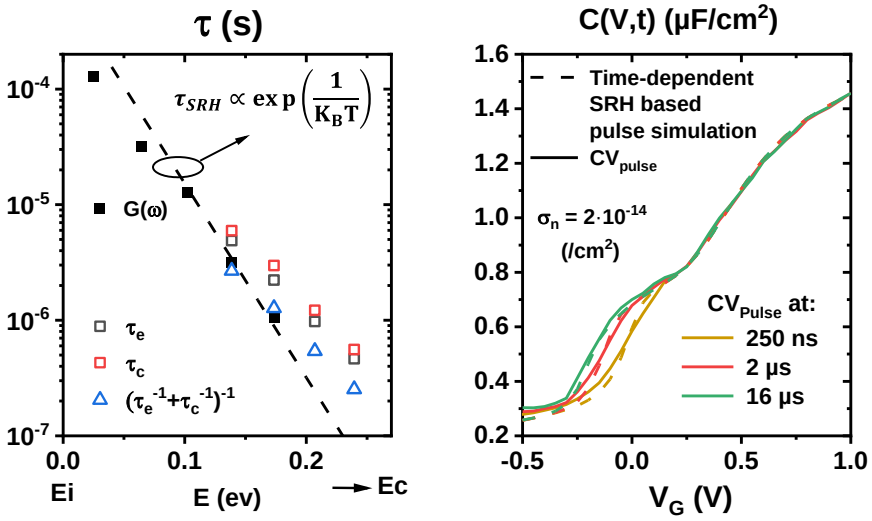


Figure 3-10(Right) Capture and emission times extracted from the fit of ΔQ_{it} with the model at different bias. They are very similar to the values obtained from $G(\omega)$. (Left) Capacitance curve obtained from CV_{pulse} compared to the SRH model at different charging moments.

$$\frac{\partial f_R^i(t)}{\partial t} = \tau_{c,i}^{-1}(t)(1 - f_R^i(t)) + \tau_{e,i}^{-1}(t) f_R^i(t) \tag{3-19}$$

$$\Delta V_{FB}(t) = \sum_{Ev}^{Ec} \frac{q \cdot Nit(E_i) \cdot f_R^i(t)}{Cox} \Delta E \tag{3-20}$$

The model perfectly captures the $Q_G(t)$ transients with a unique set of fixed parameters τ_C and τ_E; each one extracted independently at each V_{G0} from charging and discharging phases. The extracted time constants are reported in **Erreur ! Source du renvoi introuvable.**-left. They well agree with values extracted by $G(\omega)$ method, and are consistent with the SRH theory even if small discrepancies in the slope are

evidenced. This mismatch with SRH model is hardly seen when looking at the SRH modelling [6, 7] of experimental CV in **Erreur ! Source du renvoi introuvable.**-right.

3.2.3 BTI reliability using ramp pattern

Generally, BTI on MOScap consists in recording the evolution of the CV characteristics over stress time. To do that, measures-stress-measures sequences are performed with a capacimeter coupled to a signal generator. This configuration has important drawbacks since it requires a matrix to switch between the two instruments and to record the BTI drift over time (eq. 3-2). Switching from capacimeter to pulse generator takes long times >100ms, so that the BTI characterization is meaningful for stress times >1s, much larger than switching + CV measure times. To overcome this strong limitation, we propose a new fast BTI methodology based on CVramp measurement. The principle of the new technique consists in programming a single WGFMU unit to make the whole measures-stress-measures sequence i.e. applying BTI stress while by performing a fast CV during the measurement step. This can be done using a single BTI pattern applied to the gate of the MOScap ,as depicted in **Erreur ! Source du renvoi introuvable.**. These fast Stress and Measurement sequences allow to minimize the recovery time during every measurement step.

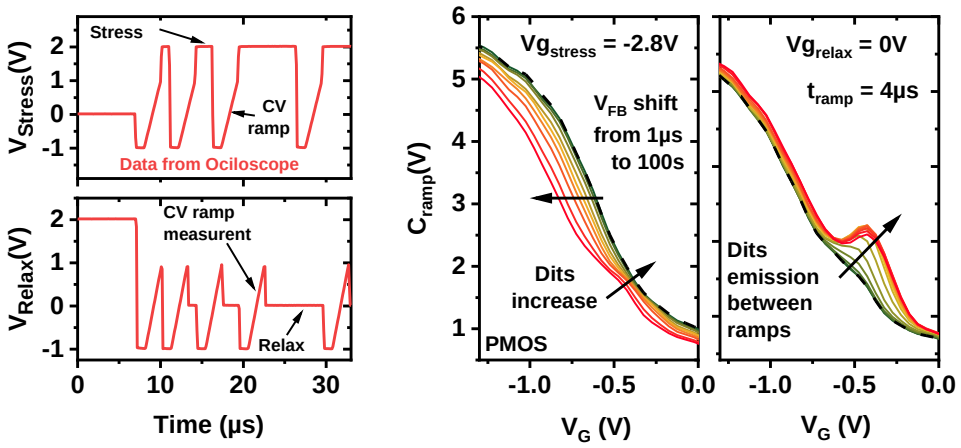


Figure 3-11(Left) Stress and Relaxation signal patterns applied for BTI reliability characterization on MOScaps. CVs are measured by a CV_{ramp} between stress intervals. (Right) Typical evolution of CV curves measured on a HfO2 p-MOScap under a -2.8V "NBTI" stress and during recovery phase at V_G=0V .

For Ultra-Fast BTI reliability characterization, the CV_{ramp} method is more attractive than CV_{pulse} as it allows even faster capacitance measurements. In this

configuration, the limitation seen in **Erreur ! Source du renvoi introuvable.**-right is not an issue, because the CV shift along V_G is constant over the stress sequence. Even if the capacitance curve is shifted from the fast CV, the variation of V_{FB} during stress is not affected. Therefore, it has minimal interference with the extraction of ΔV_{FB}.

Erreur ! Source du renvoi introuvable.-left shows the oscilloscope view from the typical patterns used for the BTI characterization [4]. On this pattern example each stress is performed at 2V followed by a 4µs CV_{ramp} ranging from -1V to 1V. The relaxation step is performed at 0V. The measured CVs over stress time of a similar pattern are presented at **Erreur ! Source du renvoi introuvable.**-Right. They exhibit both a clear V_{FB} shift due to NBTI stress and relax.

An interesting feature from these curves is that it is possible to clearly observe the D_{it} generation over NBTI stress time. It can be identified by a small bump that appears on the CV curve in the depletion regime. Although all traps are generated during the stress phase, this effect is scarcely noticeable during that phase but becomes more evident during the relax phase. The reason for this occurrence is that during the stress test, each measurement is taken before a strong bias period when all traps are captured. Since the relaxation time of the traps is too long, most of the traps remain captured during the measurement ramp. However, when the time between measurements is on the order of 100 µs, all traps can be emitted and recaptured in the subsequent measurement, producing a capacitive signal that represents the entire population of traps.

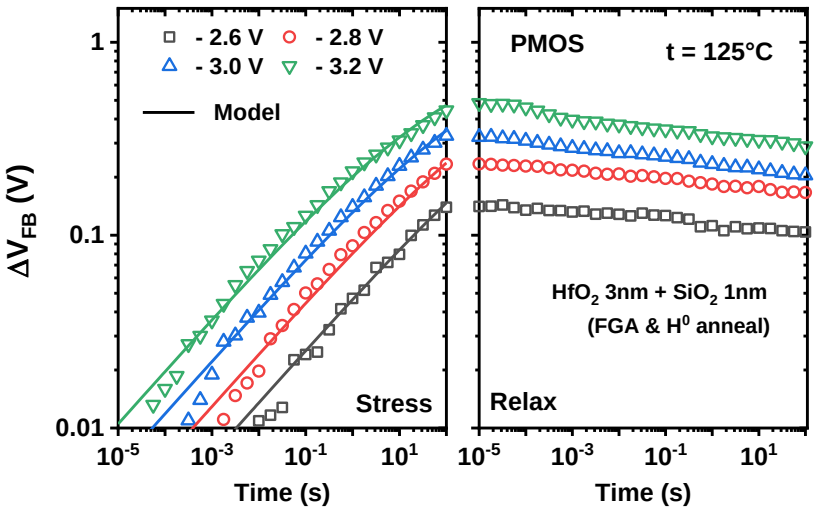


Figure 3-12 Ultra-Fast DC NBTI @T=125°C on a HfO_2 p-MOScap annealed by FGA+H0 plasma during stress and relax steps. A saturated power law model (lines) is used to extrapolate TTF in the stress phase.

The method is very effective to capture ΔV_{FB} from 10µs stress to more than 1ks, exactly as a fast BTI method applied on transistor [1]. **Erreur ! Source du renvoi introuvable.** then reports the extracted ΔV_{FB} transients during the stress and relax steps. They can be easily modelled using the power law model of eq. 3-6 to estimate the BTI device lifetime as it commonly done in MOSFETs.

$$PVG(V_G, t) = \frac{1}{sat^{-1} + A \cdot {V_G}^{-\gamma} \cdot t^{-n}} \tag{3-21}$$

This Ultra-Fast CV BTI is enough accurate to reveal the strong difference in NBTI reliability between the oxides annealed by FGA only and by FGA + plasma H^0 (**Erreur ! Source du renvoi introuvable.**) [7, 8, 10]. For this study, NBTI lifetime is found much smaller after H^0 exposure, mainly because of a change of the power law "n" exponent. This may result from a higher surface de-passivation also easily visible on the CV_{ramp} curves of **Erreur ! Source du renvoi introuvable.**-right. Indeed, in addition to ΔV_{FB}, we can obtain the evolution of interface state density due to the surface de-passivation during stress. As visible on **Erreur ! Source du renvoi introuvable.**-right, the CV curve obtained after 100 seconds of stress and relaxation steps presents a high amount of D_{it} when compared to the initial curve, revealing a de-passivation of the Si-H bonds at strong stress voltages.

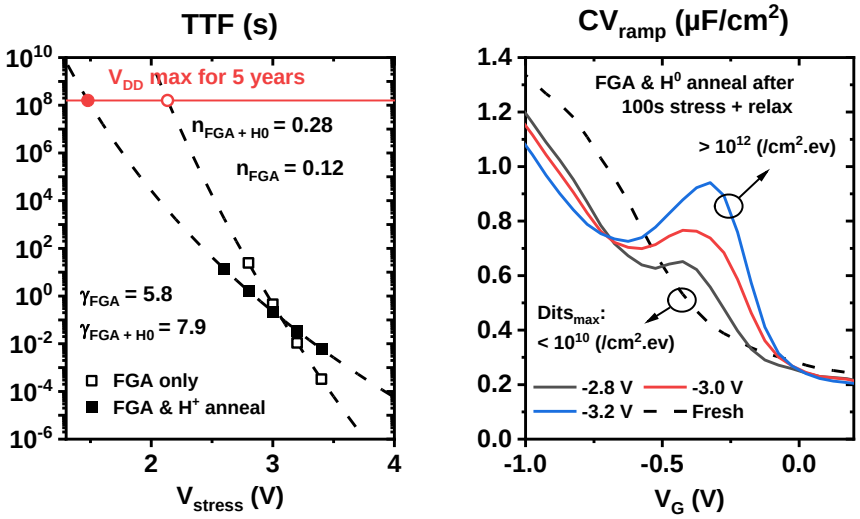


Figure 3-13(Left) BTI Time To Failure vs $V_{Gstress}$ for the two HfO2 oxides: FGA only vs FGA+H0 plasma (Right) CV_{ramp} of the H0 oxide after the stress and relaxation at various $V_{Gstress}$. Peaks are due to the increase of D_{it} during stress that do not recover.

Finally, the possibility to capture the evolution of D_{it} with stress is another key advantage of this method. In comparison, a high frequency CV method based on a capacimeter is not able to properly capture the interface states C_{it} capacitance component: it is actually difficult to distinguish the stretch from surface de-passivation from the ΔV_{FB} obtained by the capture/release of existing oxide traps. [REF]

3.3 LOW TEMPERATURE SPACER MATERIAL

3.3.1 SiCO trapping properties

3.3.1.1 T_0- CV characterization using a mercury probe setup

The first approach to understand the trapping effects in SiCO oxides was to perform full-sheet oxide characterizations. This method consists in depositing the oxide layer over a simple silicon wafer and measuring it, using the mercury drop method [Xav]. The mercury behaves like a metallic electrode above the oxide-semiconductor stack, effectively creating a MOS capacitor (MOScap). The trap characterization is thus based on classical methods described earlier like capacitance-conductance measurements. As the response of interface states occurs only in the depletion region on simple MOScap, this technique can only be used to extract the traps profile in one-half of the gap. For a complete characterization, a n and p doped substrate is required.

Erreur ! Source du renvoi introuvable. presents the capacitance measurement and the model fit for a 10nm SiCO oxide using the Hg drop. The oxide was deposited

above a 1nm thickness native SiO_2 oxide grown from the silicon substrate. Here, we can separate the traps in two categories:

- The Fast traps are midgap silicon interface defects that resonates at the capacimeter frequency range. These traps induce humps on the CV characteristics and conductance peaks in the depletion regime of GV characteristics.
- Non-resonating traps that contribute only to a frequency independent Vfb shift and to the stretch of the CV curve along V_G axis. Those traps can either be Oxide Slow border or very fast near valence band silicon interface traps.

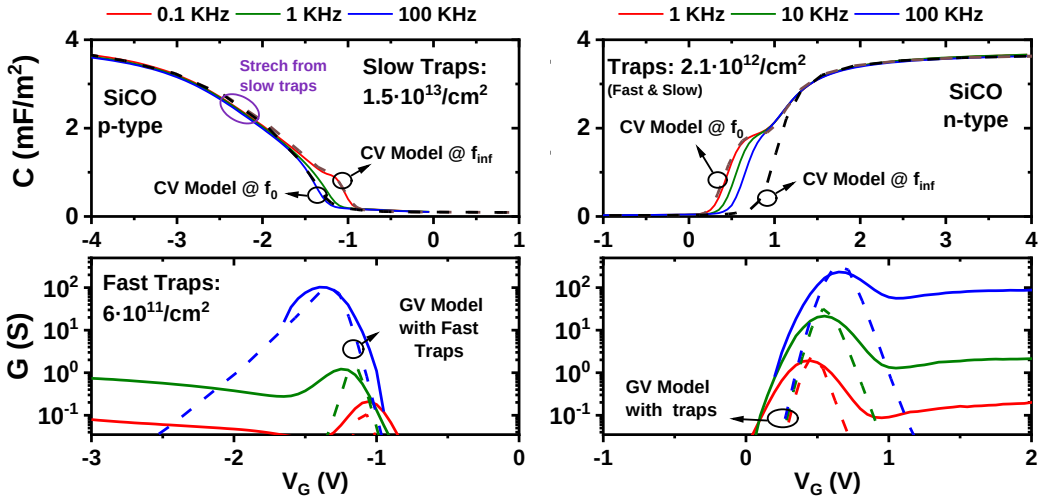


Figure 3-14 Capacitance and conductance characteristics measured on n & p - SiCO MOScaps. Fast traps are extracted using the High-Low frequency method and the results are compared to measurements using the CV simulator with this extracted trap profile. The slow trap density is extracted using Terman's method on p-type device and is found much higher than the amount of fast traps.

For the upper part of the gap, fast traps are extracted on n-type substrate using standard Low-High Frequency and conductance measurements. The extracted profile is able to well reproduce the raw CV and GV characteristics, suggesting that the density of slow oxide traps lying nearby the Si conduction band remains low in this case. However, for the lower part of the gap (p-type substrate), the extracted density of fast traps from conductance method is not sufficient to reproduce the stretch of the capacitance curve. For that, we need to include a high amount of non-resonating traps that have no frequency signature. We extract the total amount of traps using Terman's method. **Erreur ! Source du renvoi introuvable.** Figure X-left shows the extracted profile for the SiCO oxide containing both fast and non-resonating traps. We can observe that the total amount of traps is one order of magnitude higher than the one of fast traps. Therefore, these results confirm the presence of two population of traps i.e. Silicon

interface traps related to Si/SiO_2 interface defects and a second type of defects probably originated from oxyde traps in the bulk of the SiCO oxide.

To better understand the origin of the second population of traps, we performed hysteresis measurements on both devices**Erreur ! Source du renvoi introuvable.** as resented at Figure X-right. The huge shift on V_{FB} supports the hypothesis that the stretch of the CV is actually due to slow traps, that are in fact pre-existing traps of the SiCO oxide, and not to Si near Ev fast traps. These traps capture and release holes from the substrate during the sweep of the CV measurement shifting the V_{FB} of the device and stretching the curve. The hysteresis effect on N-sub devices indicates that there are also acceptors traps close to the conduction band. Those traps must be more distant from the silicon band-gap and are screened by the accumulation charge since they do not produce a stretch on the CV curve. Nonetheless, the drift on V_{fb} is dependent on the period to perform a CV sweep, the initial and final gate bias, and the direction of the sweep. The total amount of traps extracted with the classic methods will depend on how the sweep is performed and therefore are no longer reliable.

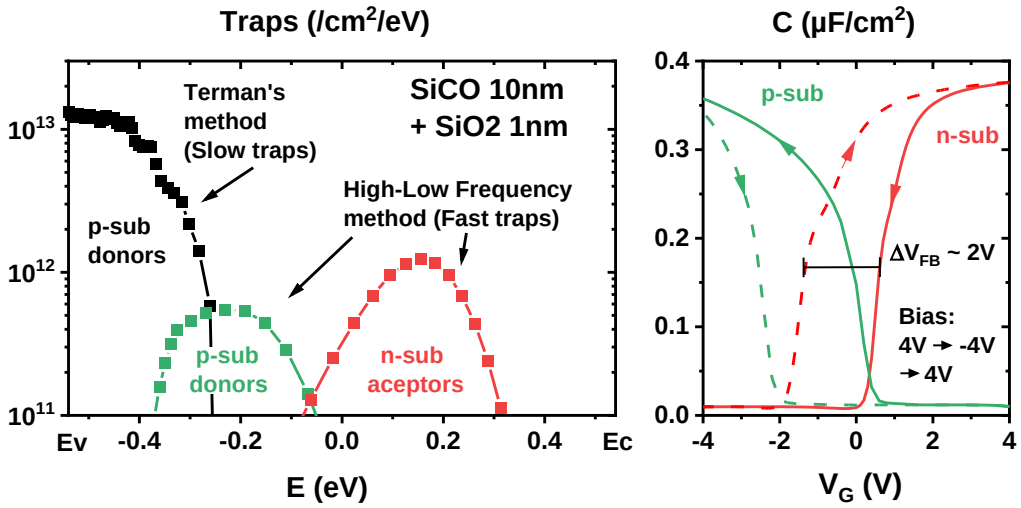


Figure 3-15(Left) Trap profile extracted from different methods on p and n type devices. The high amount of slow traps near the valence band explain the stretch on CV curve of p-sub device (Right) Hysteresis measurement of n and p type devices. Measurements are performed at 100 KHz from 4V to -4V and then back to 4V.

3.3.1.2 T_0- Dit characterization using the CV pulse technique

To go further, we characterize the quality of the Si/SiCO interface using the previously presented CV_{pulse} method. XX **Erreur ! Source du renvoi introuvable.**illustrates the variation of charge from different pulses on capture and emission conditions on the same device. We can observe distinct behaviour between

capture and emission curves. This is explained by a difference on defects capture τ_C and emission τ_E time due to a different dynamic between capture and emission probabilities, which is consistent with a SRH model for interface traps.

During the 200mV capture pulse, the Fermi potential moves closer to the valence band by a couple of kTs. This substantially increases the density of holes at the SC interface and reduces the capture time of donors traps, given by $\tau_C = c_{ps} p_s(E_{fs})(1 - f_t(E_{fs}))$ accordingly. Thus, traps are quickly captured 20 µs after the pulse. On the discharging pulse, the emission time is not governed by the number of carriers on the valence band so the emission time remains long. On this case, the traps do not react during the fast sampling period and the capacitance is more similar to the ideal CV curve.

From the variation of charge over time, we extract the CV behaviour of the SiCO oxide on different capture and emission moments. Each pulse is applied under 20µs therefore only a very small low frequency stretch is obtained on this measurement. As presented before, the slow trapping mechanism is meaningful only after milliseconds (**Erreur ! Source du renvoi introuvable.**). We can observe on **Erreur ! Source du renvoi introuvable.**-middle) a significant difference on the shape of the capacitance obtained from the fast pulsed CV to the stretched curve obtained from the capacimeter measurements. In addition, due to the different behaviour on the time constant during capture and emission, all traps react during the capture pulse while none reacts to the emission pulse. This happens because the capture time reduces with the increased density of accumulated carriers from the fast pulse, while the emission remains constant as the density of carriers reduces. Therefore, traps with emission time greater than the pulse period do not react to the CV_{ramp} independent on the gate bias.

The D_{it} profile is then extracted from the variation of charges on both capture and emission (**Erreur ! Source du renvoi introuvable.**-right). This method allows to capture a more accurate D_{it} profile extraction and to avoid misinterpretations made with the stretched CV curve obtained from the capacimeter approach. Therefore, using both CV_{pulse} and CV_{ramp} we are able to separately extract the low frequency traps responsible for the hysteresis effect from the high frequency interface state traps.

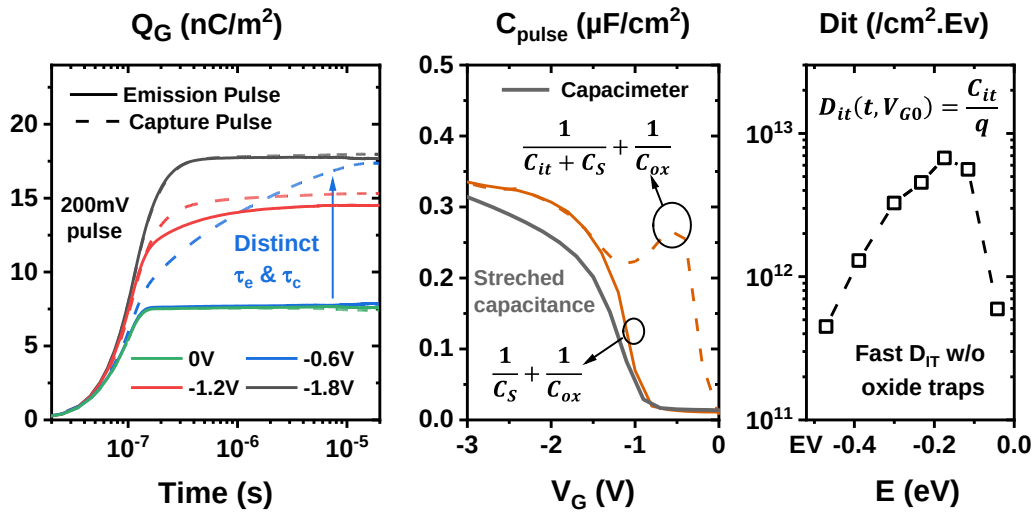


Figure 3-16(Left) Variation of Gate charge over time from CV_{pulse} on a 10nm SiCO device with 200mV pulses. Capture pulses go from positive to negative bias and Emission pulse on this opposite way (Middle) CV extracted from the charge variation from 500ns to 20µs. Within 20µs, it is possible to extract an un-stretched free of oxide traps CV unlike capacimeter (Right) D_{it} profile extracted from CV_{pulse} method considering only the fast traps. The profile is obtained using an approach similar to the High-Low frequency method.

Nonetheless, we can observe a higher amount of midgap traps from this method when compared to measurements using the mercury drop setup. This effect could be explained by two hypothesis: The relaxation time of these traps is much longer than what the capacimeter is able to measure at its lowest frequency and this effect is not apparent on the CV curves. Alternatively, the PVD Thermal evaporation deposition used to form the aluminium pads could interfere with the interface between semiconductor and oxide with may generate more traps on this setup.

3.3.1.3 Fast BTI characterization using CVramp

To investigate the effects of the slow trapping mechanism, we used the Fast CV-ramp MSM method presented early. The capacitance is first measured using an X us ramp. This CV measure is fast enough, so that this sense measurement is not modified by the slow trapping mechanism. Then a short stress is applied to the device followed by a fast measurement and so on. The shift of V_{FB} during the stress can be used to probe the amount of filled oxide traps over time. **Erreur ! Source du renvoi introuvable.** presents the Ultra-Fast BTI results obtained on the SiCO MOS capacitor at T=25°C using CV_{ramp} measurements. This characterization has not be performed with the previous setup due to limitations of the Hg contact i.e. high impedance of the mercury

drop [XX]. True MOScaps, integrating an Al PVD electrode formed by thermal evaporation, were fabricated instead.

Thanks to this setup, we can accurately characterize the trapping dynamics at different stress voltages. A huge V_{FB} instability is evidenced regardless of the stress. It starts at very short stress times before saturating after few seconds of stress. The possible reasons for this saturation are twofold:

1. The reduction of the stress oxide field with time resulting from the filling of the high density of SiCO traps. When the stress becomes too low, the generation of SiCO traps becomes negligible leading to V_{FB} saturation.
2. There is a limited amount of pre-existing traps on the oxide and no traps are been created by the stress. Thus, at long stress times, all the traps are filled, $\Delta Q_{trap} \sim 0$ and V_{FB} saturates

At this stage, it is not possible to discriminate between the two mechanisms all the more that both can advantageously combined.

The recovery dynamics of the slow oxide traps is also investigated in **Erreur ! Source du renvoi introuvable.**. It appears that V_{FB} can be fully recovered if enough time is given to traps to relax. This last result finally suggests that the SiCO oxide actually behaves as a memory-like device, with a large programming window seen on V_{FB}. This large instability can be a serious concern when SiCO material is used as spacer oxide, but may be interesting for another applications like charge trapped memory devices.

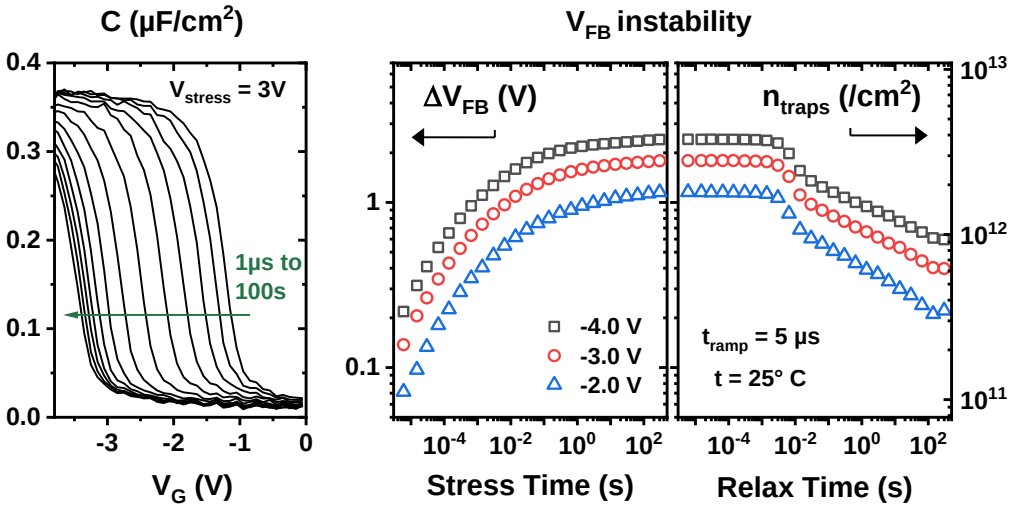


Figure 3-17 Ultra-Fast MSM on a 10nm SiCO p-MOScap using CV_{ramp} measurements. (Left) Evolution of the CV characteristics over stress time at $V_{Gstress}$=3V (Right) Variation of V_{FB} for different bias conditions over stress and relaxation time. V_{FB} largely shifts because of the charging/discharging of pre-existing recoverable oxide traps. Common capacimeter MSM method cannot characterize properly this shift because the trapping occurs during the CV measure itself at low voltage.

The same study was performed on n-type substrate with the same oxide stack (**Erreur ! Source du renvoi introuvable.** left) using equivalent stress conditions. We can see that the shift of V_{FB} on this case is smaller than for the p-type substrate. From the shift on V_{fb} and the stress bias it is possible to estimate the total amount of traps that are captured at the end of the stress as function of the surface potential. Finally, **Erreur ! Source du renvoi introuvable.** right presents the traps extracted using this technique to the previous profile extracted using Terman's method. Two features are highlighted:

1. The density of traps from UF-MSM setup is higher than the one from Terman's method. The difference may lie in the fact that a noticeable amount of traps cannot be captured by Terman's method due to recovery effects inherent to the CV measurement using a capacimeter and matrix switching.

2. It confirms that the energy states of oxide traps on a n-substrate lie above the conduction band of Si. Because of their energy position, these oxide traps will can be filled only for V_G in the strong accumulation regime, and thereby do no stretch the CV characteristic unlike in p-type substrates.

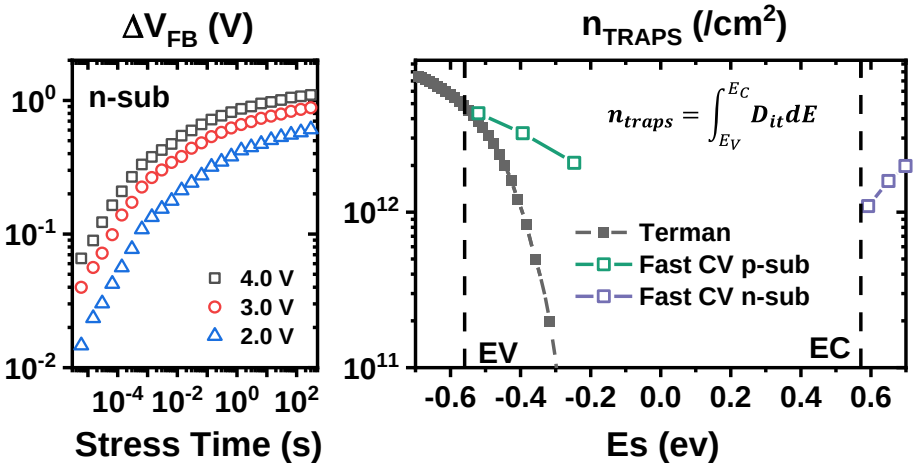


Figure 3-18 Ultra-Fast MSM on a 10nm SiCO n-MOScap using CV_{ramp} measurements. (Left) Comparison of the total amount of donors and acceptor oxide traps over the gap using the MSM setup against the profile obtained from Terman's method with the capacimeter setup.

3.3.2 Annealing effects on SiCO oxides

From the previous experiments, we can conclude that the MOScaps based on an as-deposited low-k SiCO material contain a huge amount of traps. If the density of interface defects is more actually related to the quality to the native oxide between the SiCO and the Si, the large amount of slow oxide traps, that are not present on thermal SiO_2 oxides, is intrinsic to the porous SiCO material itself. Moreover, it is clear that this trapping effect will have detrimental effects on MOSFET performance if the oxide is used as spacer. Due to the proximity to the channel and access regions, the charging and discharging of the oxide traps can impact its FoMs like V_T, DIBL and access resistance, as it will be further demonstrated in the following session. It is therefore crucial to reduce the amount of defects if we want to use this material as spacer.

Actually the SiCO oxide traps look like oxygen vacancies defects that are acceptors traps formed by a missing oxygen atom in a Si-O-Si bond on non-stoichiometry dielectrics. They are frequently responsible for hysteresis effects on gate oxides [ref] and are present on oxides grown from SiC substrates [ref]. During the deposition process, the carbon atoms of the SiCO oxide, that are more electronegative than Si, can get the oxygen atom from Si-O-Si bonds, creating oxygen vacancy defects [ref]. To reduce this defect number, some anneals like Forming Gas, Dry O_2, N_2O, NO and CO_2 anneals were proposed in the literature [ref]. The Forming Gas Anneal (FGA) is an interesting option because it is a well-known anneal frequently used by the semiconductor industry: in CMOS technology, it is generally performed at the end of the front-end integration process to passivate oxide and interface defects.

We therefore performed a 40 minutes 350°C FGA anneal on the same SiCO oxides than the ones studied in FIG. The results are presented on **Erreur ! Source du renvoi introuvable.**. We can clearly observe a drastic improvement on both trapping mechanisms after the FGA. The density of oxide traps is reduced by a factor 5. The CV curve still is slightly stretched when compared to an ideal CV, proving that the quality of the SiCO yet remains lower than the one of a standard thermal SiO_2 oxide. However, the requirements for a spacer oxide are less than for a gate oxide, and such a density may be finally acceptable. The reason for that, is that a same amount of defects in the spacer oxide has considerably less impact on the device FoMs than in a gate insulator. Similarly, the interface state density is reduced, but the improvement was not as effective as we would expect from this anneal. It indicates that not all defects are passivated with this process, and that maybe longer or different temperatures anneals are suitable for even better results.

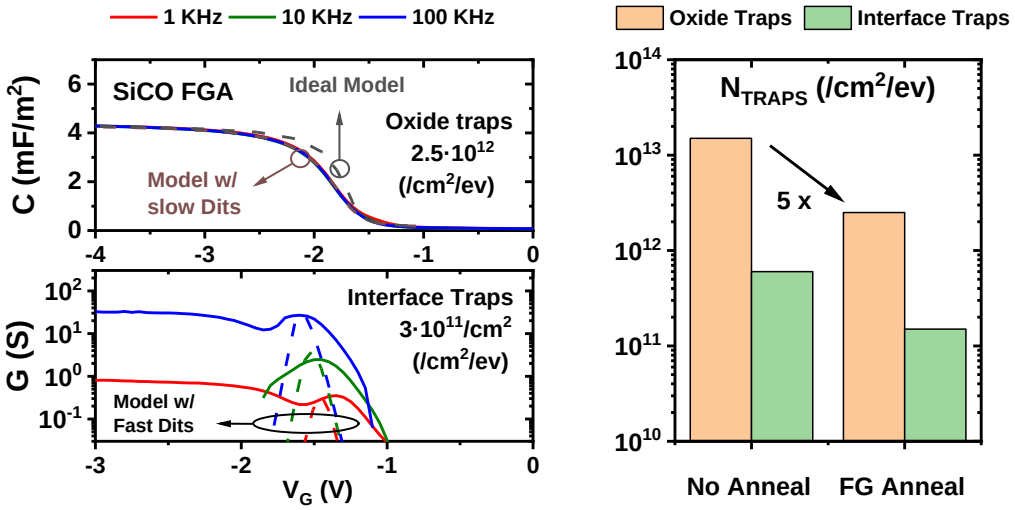


Figure 3-19(Left) Capacitance and conductance measurements of SiCO MOScap with a p-type substrate after a FG anneal. Fast interface traps are extracted using the High-Low frequency method while slow oxide traps are extracted using Terman's method. (Right) comparison between oxide and interface traps before and after the forming gas anneal. EOT of the oxide stack is does not change between anneals

Another important aspect is to understand the effects of a high thermal budget process on the oxide in order to evaluate the suitability of this low-k material for other technologies than LTB. With this motivation, we performed a 30s 1080°C Spike anneal on the same low temperature SiCO oxide stack. This anneal mimics the process used in HTB technologies for the activation of source and drain dopants.

Erreur ! Source du renvoi introuvable.-left shows the hysteresis measured with a capacimeter for both reference and annealed devices. From the big reduction on

the shift of V_{fb}, we could conclude that the slow trapping mechanism is drastically reduced after the high temperature anneal. This effect has been also observed by [ref] when the oxide deposition temperature is increased. However, a fast CV_{ramp} MSM characterization on the same device reveals that the shift on V_{FB} after 100s stress is improved only by 500mV. The big difference between both methods is explained by the recovering dynamics after stress. Indeed the shift of V_{FB} after 100s relaxation for the reference device is still between 200mV and 600mV, while it is almost completely recovered for the annealed device. This again demonstrate the importance of the fast MSM method for a right description of the device trapping behaviour.

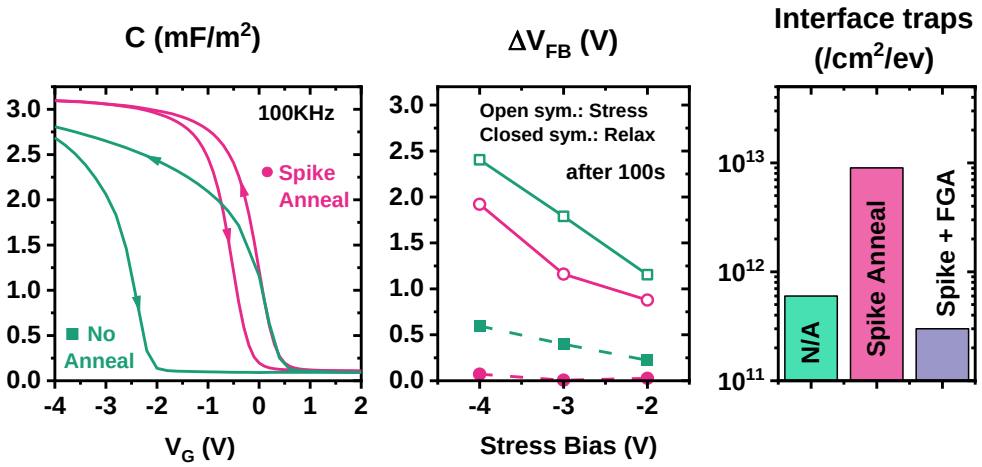


Figure 3-20 Comparison of FoM between SiCO oxide stack MOScap before and after a 1050°C Spike anneal (left) CV hysteresis measurement performed between 4V and -4V at 100 kHz with a capacimeter. (Middle) Fast CV_{ramp} MSM characterization of flat band shift from different bias conditions after 100s stress ($V_{relax} = 0$) (open symbols) and relax (closed symbols) (right) Effect of spike anneal and forming gas on the density of Interface traps extracted from G(ω) curves.

Finally, **Erreur ! Source du renvoi introuvable.**-right reports the interface trap density for all the splits. The spike anneal alone is responsible for a drastic increase of the Dit density. However, a subsequent FGA can easily cancel the degradation induced by the HT anneal. Therefore, by combining Spike and FGA anneals, it is possible to "tune" the trapping properties of the SiCO material. This may be useful if we want to use SiCO for other applications than MOSFET spacer.

3.3.3 Conclusion on SiCO trapping properties

In this part we studied the trapping properties of SiCO. We evidence two population of traps presents on SiCO oxides: A first population related to interface traps probably related to the interface between silicon and the native oxide already

present before the SiCO deposition. A second population of slow reaction traps related to pre-existing defects inside the oxide that resemble oxygen vacancies defects.

So far we focus on SiCO material as the replacement of SiN for MOSFET spacer application. The requirements were then to reduce as much as possible the density of traps within the oxide. However, for other applications, it could be interesting to maximize this trap density instead. This is the case if want to use SiCO as a trapping layer for non-volatile SONOS memories in replacement of SiN. Another great potential for SiCO could be to integrate it as a trap rich layer in RF substrates. Indeed nowadays , where a high amount of traps are useful to reduce electromagnetic coupling with the substrate this material could be used as a sink of carriers to fix the potential at the substrate in a depletion regime. With a pinned potential, it is possible to avoid the formation of inversion or accumulation layer under the BOX responsible to an increase on losses and reduction of the quality factor of inductors and transmission lines.

3.4 EFFECTS OF SPACER CHARGES ON DEVICE PERFORMANCE

3.4.1 Impact on transistor IV performance

To understand the effects of spacer interface charges on device performance, we performed TCAD simulations of LTB MOSFET transistors with SiCO spacers. Charges were added to the interface between spacers and substrate to identify the possible effects as illustrated at **Erreur ! Source du renvoi introuvable.**-left.

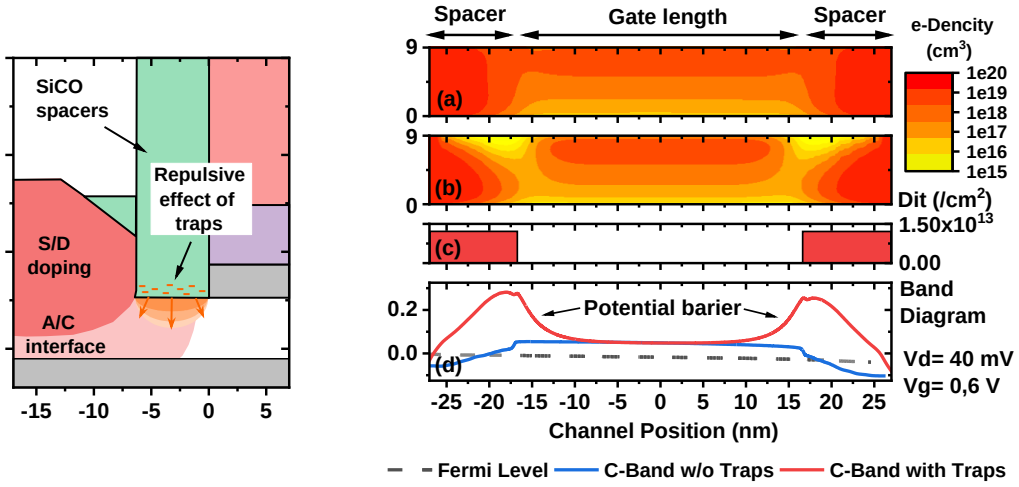


Figure 3-21(Left) Sketch illustrating the localization of the charged acceptor traps underneath the spacers. (Right) Electron density in the channel and spacer regions of a 30nm NMOS transistor without (a) and with (b) spacer interface charges ad $V_D = 0V$ and $V_G = 1V$. (c) Dit profile over the channel length. (d) Band Diagram for both scenarios.

The simulations were performed using a constant D_{IT} profile with the same density as the one extracted from the previous session. The defects closer to the valence band are considered as donor states (+/0) while those closer to the conduction band are considered as acceptors (-/0). When a trap captures a free charge, it will produce an electrical field, that can either repel or attract carriers on the substrate depending on charge polarity. For NMOS devices, the majority carriers are electrons, so that, in the spacer region, charged donor traps will attract more electrons close to the interface while charged acceptors will repel them. Instead, for PMOS devices, donor traps will repel and acceptors will attract majority carriers. However, in the true device operating conditions, only trap repulsive effects will be observed for the two types of transistors. Basically this results from the fact that, due to the high doping density in this region combined to the polarity applied to the gate, the Fermi level at the spacer interface always lies very close to the conduction and valence band for NMOS and PMOS respectively.

The doping profile of the S/D junctions is another important aspect of the simulation. The lesser is the dopant density under the spacers, the stronger will be the repulsive effect of the traps. For high-doped overlapped junctions, the electrical field induced by the SiCO charges can be easily screened out by the high density of majority carriers. However, for underlapped devices, the spacer charges can easily deplete a large portion of the silicon substrate in the junction region, that affects the final device performance. Therefore, in the following, a worst-case study is performed on a non-optimized underlapped device. To represent it on the simulation, we used a simple Gaussian profile that matches the CV behaviour of a short channel device.

Erreur ! Source du renvoi introuvable. illustrates the TCAD results on the carrier density and substrate band diagram for this device. We compare the simulation of a NMOS transistor without (**Erreur ! Source du renvoi introuvable.**-left-top) and with (**Erreur ! Source du renvoi introuvable.**-left-bottom) charges under the spacers. Indeed, the electrical field induced by the trapped charges repels a considerable part of negative carriers on the channel, creating a depleted region in the vicinity of the spacer. The presence of this depleted region is crucial for device performance because it will directly affect access resistance and short channel effects.

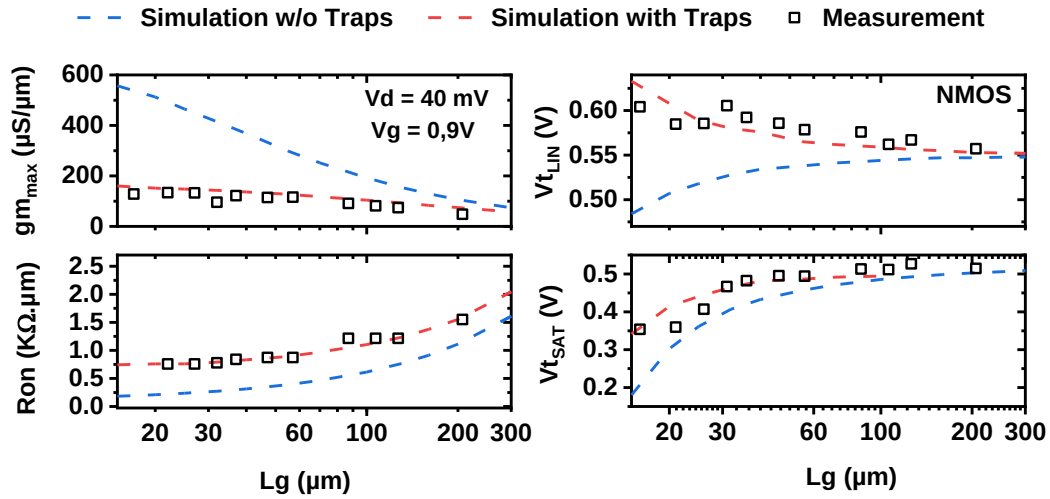


Figure 3-22(left-top) Measured and simulated maximum value of linear gm as a function of channel length with (red) and without (gray) spacer interface charges. (left-bottom) Variation of R_{on} with L for the same scenarios, the vertical deviation of R_{on} indicates an increase of the access resistance. (right) Measured and simulated Threshold voltage with (Red) and without (Gray) spacer interface charges. A great impact of the charges is visible on NMOS (right-top) while there is a negligible impact on PMOS (right-bottom) because of the overlapped channel.

In addition, IV simulations were performed on devices with different channel lengths, and the key FoMs were extracted for both cases. The results are also compared to the parameters extracted from the measurements of the real device. **Erreur ! Source du renvoi introuvable.**-left shows the impact of spacer charges on the NMOS linear gain (g_m) and channel resistance R_{on}. The depleted region created under the spacers increases significantly the access resistance, thereby degrading the total resistance R_{on} and the linear gain. This increase of the access resistance is visible whatever the channel length. The simulation without interface traps leads to much higher values of g_m and R_{on} than the real ones obtained from measurements. However, when traps are added to the spacer, the variation of the measured parameters over channel length is perfectly reproduced by our TCAD simulations.

The effects on linear and saturation threshold voltage are also studied. **Erreur ! Source du renvoi introuvable.**-left-right shows the variation on linear and saturation threshold voltage induced by spacer charges of NMOS and PMOS devices. The measured NMOS linear V_T weirdly increases for short channel devices in contrast to the expected trend from short channel effects behaviour. Indeed, the V_T extracted from simulation without traps actually decreases for smaller channel lengths. Again, when charges are added to the simulation, the results perfectly captures the variation of both measured linear and saturation V_T over channel length.

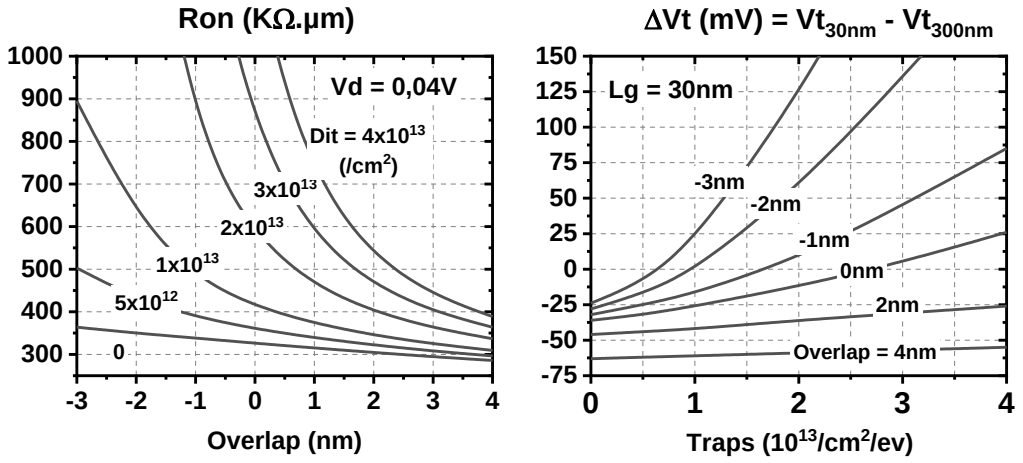


Figure 3-23 TCAD simulation of 30nm devices (Left) R_{on} versus overlap position for several D_{IT} densities. (Right) linear $V_t(L_G = 30nm) - V_t(L_G = 300nm)$ versus D_{IT} for different overlap positions. The effect of D_{IT} is clearly enhanced on underlapped devices.

Nevertheless, to better understand how spacer charges interact with channel dopants, TCAD simulations were performed for various densities of interface traps

and junction positions (see **Erreur ! Source du renvoi introuvable.**). The spacer charges modify R_{on} whatever the doping condition. However, the resistance degradation is greatly attenuated on overlapped transistors. A similar behavior is evidenced on V_t. The stability of $V_t$with respect to a long channel device $V_t(L_G = 30nm) - V_t(L_G = 300nm)$ strongly depends on the overlap nature: V_t is much more stable for overlapped transistors than in their underlapped counterparts. The behaviour of a transistor without charges is also reported for comparison. It confirms that the increase of linear V_t and R_{ON} measured in our SiCO transistors cannot be explained only by SCE independent of the overlap.

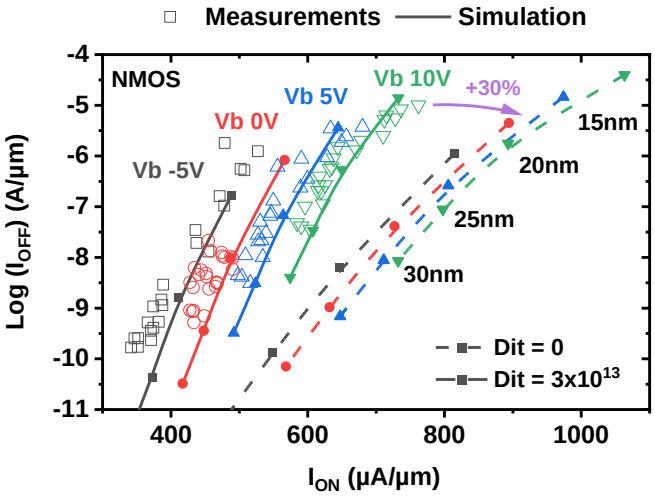


Figure 3-24 Measured (symbol) and simulated (line-symbol) I_{ON}/I_{OFF}. characteristics of the NMOS transistor for different back bias conditions. The dashed curves represent the scenario without trapped charges under the spacers.

Finally, we explore the effect of the spacer traps on the relationship between saturation on-state and off-state current by simulation, and compare it with the real one from measurements. **Erreur ! Source du renvoi introuvable.** presents the impact of the traps on I_{ON}/I_{OFF}. The different back bias (V_B) conditions allow probing the effect of traps when the inversion layer is created further from the gate oxide interface. For higher positive values of V_B, the charge centroid of the channel inversion layer becomes closer to the box interface that is almost free of interfacial traps. The impact of spacer charges must be therefore lesser on Ron if the degradation of access resistance comes from the interface with the spacers. Indeed, the simulation case with charges present the same trend and fit the measured values over different values of V_B. In addition, the substantial increase of Ion at almost same I_{OFF} indicates a reduction of the access resistance for high values of V_B. Finally, the simulation of the 'trap free'

transistors show that a huge performance gain (+30%) can be expected for the same device without spacer interface traps.

3.4.2 Impact on transistor BTI performance

To explore the effects of the spacer charges on reliability, PBTI measurements were performed on both NMOS SiCO and SiN spacers based transistors processed at low temperature. The measurements were performed using an ultra-fast method with µs resolution [9]. Fig. x reports the V_T drift with time and the corresponding time to failure for both kind of devices. Both SiCO and SiN spacer transistors meet the 5 years criterion at $V_{DD} + 10\%$. However, no clear correlation can be found between the BTI degradation and the quality of the spacers. This suggests that BTI, even in these low temperature short channel devices, is mainly driven by the quality of the gate oxide rather than by the one of the spacers. Yet, as presented on the previous chapter, the HC degradation instead is highly dependent on the overlap position. It is then expected to be largely influenced by the spacer charges in the case of underlapped devices.

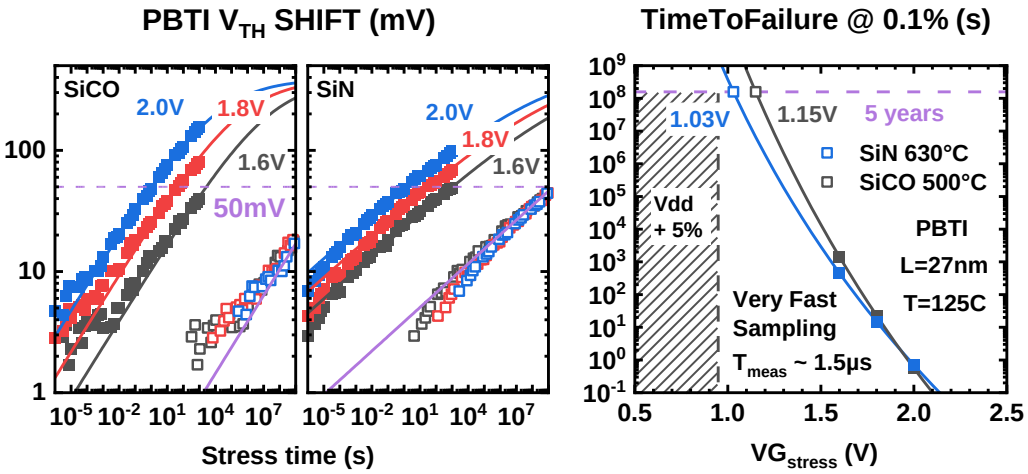


Figure 3-25 (left) Ultra-Fast BTI performed on 27nm NMOS transistors with SiCO and SiN spacer at 3 stress conditions (symbols). A power law model (lines) is used to extrapolate TTF at V_{DD} + 10% (open symbols). (right) BTI Time To Failure vs $V_{G\,stress}$. VG@5years is found higher in SiCO than in SiN based devices because of a higher voltage acceleration factor. Yet, both transistors meet the years requirements for this technology $V_{DD} = 0.9V$

3.5 CONCLUSION

We proposed two new fast capacitance measurements techniques named CV_{pulse} & CV_{ramp} to characterize traps and ageing of MOS capacitors. The CV_{pulse}, is suitable for trap spectroscopy it allows an efficient reconstruction of the CV at different filling rates of interface traps on devices containing $D_{it} > 10^{11}/cm^2eV$. On the other hand, the CV_{ramp} method is useful for Ultra-Fast BTI as it allows monitoring the full dynamics of trapping (µs to ks stress) exactly as in MOSFET devices, without being impacted by undesirable effects inherent to long measurement time of a capacimeter. Both methods can be used to perform the same deep oxide characterizations than in MOSFETs using simple MOScap devices. This is obviously beneficial for several lower reasons: lower complexity of the process, lower cost and shorter cycling for gate stack optimization.

The new characterization techniques were used to characterize the trapping mechanism on the SiCO material used for spacers of LETI LTB MOSFETs. CV_{ramp} allows to characterize the low frequency trapping mechanism while the short pulses CV_{pulse} can be used to perform an accurate extraction of the interface state profile. 2 types of defects were then identified in SiCO oxides

- Fast Si interface traps which density depends on the quality of the native oxide
- Slow deep defects distributed in the bulk of the SiCO oxide.

The study was pushed forward to understand the effects of forming gas and Spike annealing on SiCO trap passivation and de-passivation. We concluded that a forming gas anneal was effective to passivate a large amount of both oxide and interface traps. On the other hand, a High Temperature annealing de-passivates the Si interface generating a large density of interface traps while reducing the deep oxide trap density. This increase in interface trap density by the thermal anneal can be partially recovered with a subsequent forming Gas anneal.

Finally, the effects of interface traps in the spacer region on performance and reliability of 30nm FDSOI low temperature transistors has been explored by means of CV measurements and TCAD simulations. It is demonstrated that the presence of spacer charges induces the formation of a depleted zone below the spacers responsible for increasing access resistance and abnormal V_T variation with device scaling. We

also demonstrated that the presence of dopants near the defects attenuates the negative effects of the interface charges. Therefore, this issue is detrimental only for underlapped channels. TCAD simulations of the transistors containing charges under the spacers correlate to the measurements and allows to predict the improvement of performance expected for a device without charges. Finally, in spite of the device performance, the PBTI reliability is not affected by the quality of the material in the spacers, even for very short channel devices.

Most of the results from this chapter were presented at the International Reliability Physics Symposium IRPS-2021 in and IRPS -2022.

3.6 REFERENCES

[1] L. Brunet et al., "First demonstration of a CMOS over CMOS 3D VLSI CoolCube™ integration on 300mm wafers," 2016 IEEE Symposium on VLSI Technology, Honolulu, HI, 2016, pp. 1-2, doi: 10.1109/VLSIT.2016.7573428.

[2] P. Batude et al., "3D sequential integration opportunities and technology optimization," IEEE International Interconnect Technology Conference, San Jose, CA, 2014, pp. 373-376, doi: 10.1109/IITC.2014.6831837.

[3] Z. Wu et al., "Accelerated Capture and Emission (ACE) Measurement Pattern for Efficient BTI Characterization and Modeling," 2019 IEEE International Reliability Physics Symposium (IRPS), Monterey, CA, USA, 2019, pp. 1-7, doi: 10.1109/IRPS.2019.8720541.

[4] A. Tsiara et al., "Performance and Reliability of a Fully Integrated 3D Sequential Technology," 2018 IEEE Symposium on VLSI Technology, Honolulu, HI, 2018, pp. 75-76.

[5] J. Franco et al., "BTI Reliability Improvement Strategies in Low Thermal Budget Gate Stacks for 3D Sequential Integration," 2018 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2018, pp. 34.2.1-34.2.4.

[6] X. Garros et al., "RF Performance of a Fully Integrated 3D Sequential Technology," 2019 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2019, pp. 25.1.1-25.1.4.

[7] C. Fenouillet-Beranger et al., "First demonstration of low temperature (≤ 500°C) CMOS devices featuring functional RO and SRAM bitcells toward 3D VLSI integration," 2020 IEEE Symposium on VLSI Technology, Honolulu, HI, USA, 2020, pp. 1-2.

[8] E. H. Nicollian and J. R. Brews, "MOS (Metal Oxide Semiconductor) Physics and Technology," John Wiley & Sons, New York, 1982.

[9] X. Garros et al., "PBTI mechanisms in La containing Hf-based oxides assessed by very Fast IV measurements," 2010 International Electron Devices Meeting, San Francisco, CA, 2010, pp. 4.6.1-4.6.4, doi: 10.1109/IEDM.2010.5703297.

[10] X. Garros et al., "PBTI mechanisms in La containing Hf-based oxides assessed by very Fast IV measurements," 2010 International Electron Devices Meeting, 2010, pp. 4.6.1-4.6.4, doi: 10.1109/IEDM.2010.5703297.

[11] J.H. Stathis, S. Mahapatra, T. Grasser, Controversial issues in negative bias temperature instability, Microelectron. Reliab. 81 (2018) 244–251.

[12] P. R. Shrestha, K. P. Cheung, J. P. Campbell, J. T. Ryan and H. Baumgart, "Accurate Fast Capacitance Measurements for Reliable Device Characterization," in IEEE Transactions on Electron Devices, vol. 61, no. 7, pp. 2509-2514, July 2014, doi: 10.1109/TED.2014.2325674.

[13] A. G. Viey et al., "Study on the difference between ID(VG) and C(VG) pBTI shifts in GaN-on-Si E-mode MOSc-HEMT," 2021 IEEE International Reliability Physics Symposium (IRPS), 2021, pp. 1-8, doi: 10.1109/IRPS46558.2021.9405221.

[14] D. R. Aguado, B. Govoreanu, W. D. Zhang, M. Jurczak, K. De Meyer and J. Van Houdt, "A Novel Trapping/Detrapping Model for Defect Profiling in High- k Materials Using the Two-Pulse Capacitance–Voltage Technique," in IEEE Transactions on Electron Devices, vol. 57, no. 10, pp. 2726-2735, Oct. 2010, doi: 10.1109/TED.2010.2063292.

[15] E. H. Nicollian and J. R. Brews, "MOS (Metal Oxide Semiconductor) Physics and Technology," John Wiley & Sons, New York, 1982.

[16] X. Garros et al., "Process damages in HfO/sub 2//TiN stacks: the key role of H/sup 0/ and H/sub 2/ anneals," IEEE InternationalElectron Devices Meeting, 2005. IEDM Technical Digest., 2005, pp. 4 pp.-194, doi: 10.1109/IEDM.2005.1609303.

[17] X. Garros et al., "Interface states in HfO/sub 2/ stacks with metal gate: nature, passivation, generation [MOS devices]," 2005 IEEE International Reliability Physics Symposium, 2005. Proceedings. 43rd Annual., 2005, pp. 55-60, doi: 10.1109/RELPHY.2005.1493062.

[18] X. Garros, J. Mitard, C. Leroux, G. Reimbold and F. Boulanger, "In Depth Analysis of VT Instabilities in HFO2 Technologies by Charge Pumping Measurements and Electrical Modeling," 2007 IEEE International Reliability Physics Symposium Proceedings. 45th Annual, 2007, pp. 61-66, doi: 10.1109/RELPHY.2007.369869.

[19] J. Franco et al., "Atomic Hydrogen Exposure to Enable High-Quality Low-Temperature SiO2 with Excellent pMOS NBTI Reliability Compatible with 3D Sequential Tier Stacking," 2020 IEEE International Electron Devices Meeting (IEDM), 2020, pp. 31.2.1-31.2.4, doi: 10.1109/IEDM13553.2020.9372054.

Chapter 4: High Frequency performance of Low Thermal Budget devices

4.1 INTRODUCTION

The use of silicon CMOS millimeter-wave (mm-wave) circuits in automotive radars operating at 77GHz [Rekha Yadav], silicon mm-wave backhaul transceivers at 57–86 GHz and 60-GHz wireless high-definition multimedia interface (HDMI) systems [Eldad Perahia Eldad Perahia] have been relatively widespread in the recent years [Sorin P. Voinigescu]. The acceptance of silicon technologies at mm-wave frequencies makes the industry is now convinced of their feasibility and benefits [Voinigescu et al] with respect to III-V materials.

In contrast, for digital applications, not all high frequency building blocks have benefited from technology scaling. Despite the improvement on transconductance, power amplifier (PA), voltage-controlled oscillator (VCO) suffer from low voltage swing and worse noise performance [book] from scaled MOSFET transistors. In addition, RF performance of Si technologies peaks at 40 nm and 22 nm CMOS nodes. Shorter channel length FinFET and Gate all around (GAA) devices present lower performance due to higher parasitic gate capacitance and resistance. Therefore, as power consumption, density and cost per standard cell of digital devices are reduced with scaling, it would be relevant to integrate RF and digital blocks on separated substrates with different process nodes for SoC applications. In this way, digital blocks would benefit from the high performance but expensive advanced node process, while the high frequency and analog components could be integrated with a low-cost relaxed process.

3D Sequential Integration (3DSI) opens the possibility for hybrid integration of RF, analog and digital functions on several front-end tiers sequentially fabricated (**Erreur ! Source du renvoi introuvable.** and **Erreur ! Source du renvoi introuvable.**). The stack of active device layers on top of each other with small 3D contact pitch (<100 nm), allows a high transistor density and reduced parasitic and interconnect length [refs]. Such key features offer great potential for 5G mm-wave

applications, which require a high density of antennas nearest to power and low noise amplifiers and analog ADC circuits [5].

On this context, the 28-nm FDSOI CMOS technology, one of the most advanced commercial SOI node, is a good candidate for the top tier of high frequency systems. It provides low leakage and back biasing capabilities for low power digital and mixed-signal (MS)/RF applications. Furthermore, despite their intrinsic limitations for high frequency applications, silicon MOSFET transistors still provide the lowest fabrication cost with high FoM.

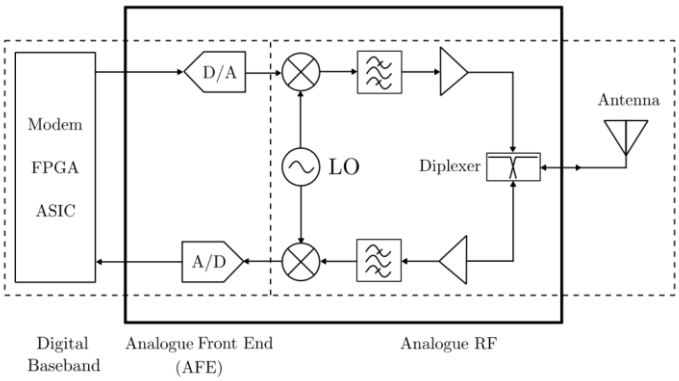


Figure 4-1 Generic block diagram of a wireless transceiver [State-of-the-Art Millimetre-Wave Silicon Transceivers and Systems-on-Chip]

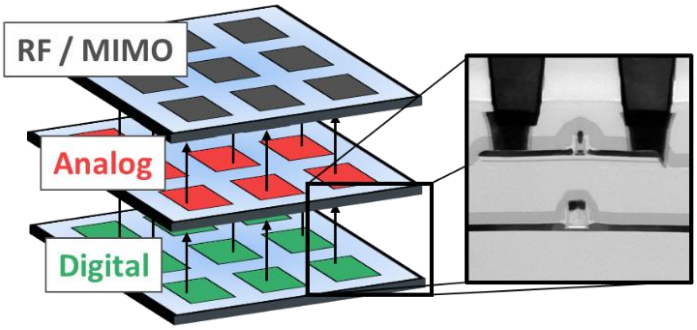


Figure 4-2 (Left) Architecture of a system integrating digital, analog and RF circuits on separated layers. The high-density interconnection can be useful for parallelization of decoding functions on MIMO systems. (Right) Transmission Electron Microscopy (TEM) picture of two stacked MOS transistors [REF].

Likewise, the heterogeneous sequential integration of RF systems can be useful for the development of silicon-based quantum computers. Indeed, gate reflectometry, which is a promising method for the spin readout of CMOS quantum dots, requires the integration of RF circuits [3]. So far die-to-wafer hybrid bonding with 20-µm interconnections are already under study for cryogenic applications [4]. 3DSI may be

the next step since, thanks to its small footprint and the high density of interconnects, it is very suitable to reduce the wiring that serves to address and read quantum bits.

This chapter proposes an extensive study of high frequency performance of low temperature devices developed for 3DSI CMOS technology in order to know if 3DSI can meet requirements for some RF applications. For this purpose, the effects of low temperature fabrication steps on F_T and F_{MAx} are studied as they are efficient FoMs to link the RF electrical response to key technological parameters.

The first part of this chapter describes a new technique for the extraction of device effective mobility from high frequency measurements, and gives and one application for the study of Low thermal budget (LTB) transistors. Then, the impact of parasitic capacitance SiCO spacers and access resistances on the device performance FOMs is addressed . The effect of low temperature gate annealing on the gate resistance, and its impact on device RF performance, is also investigated. Finally, the performance of LTB devices fabricated by LETI are compared to the state of the art silicon MOSFET transistors found elsewhere.

RF characterization is performed on two different FD-SOI technologies developed by LETI, the 28-nm and a simplified 45-nm node. Both processes have similar fabrication steps. The only difference lies only on the use of two metal layers and the lack of a top substrate contact for the 45-nm node. High temperature devices fabricated with a classic process flow are used as benchmark for both technology nodes. The measured devices are multi-gate finger transistors integrated with OPEN and SHORT de-embedding structures for parasitic correction. “S parameters” measurements have been performed at wafer level and room temperature with GSG RF probes using the Keysight N5245B PNA-X Microwave Network Analyzer that works in the frequency range of 500 MHz-50 GHz. The small circuit models are extracted using the same method as [ref].

4.2 RF FIGURES OF MERIT

The figures of merit (FoM) of MOSFET transistors designed for high frequency applications are intrinsically different from those of analog and digital devices. While digital and analog FoMs are highly related to the saturated drain current, leakage, linearity and voltage gain in low frequencies. In particular, the devices capacitance and gate resistance have a greater contribution on high frequency FoM than in analog and digital transistors. **Erreur ! Source du renvoi introuvable.** recalls the RF small-signal equivalent circuit of a MOSFET. Equations (4-1) and (4-2) describe the relation between f_t and f_{max} with the key parameters derived from the electrical model proposed in [14].

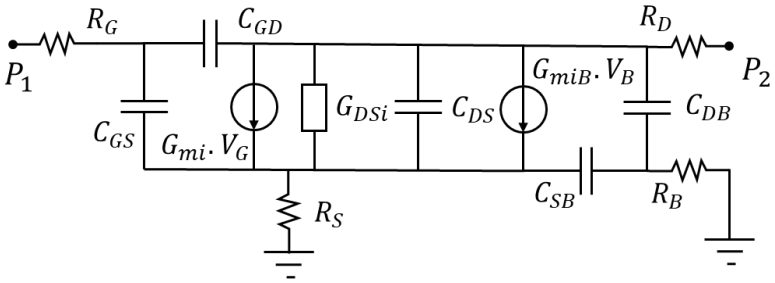


Figure 4-3 RF small-signal equivalent circuit of a two-port common source MOSFET where the substrate and the source are connected to the ground. Port 1 (P1) is connected to the gate while the port 2 (P2) is connected to the drain.

The FoMs f_T and f_{max} can be expressed as the function of the classical electrical device parameters as given by eq. 4-1 and eq. 4-2.

$$f_T = \frac{g_M}{2\pi \cdot C_{GG}} \tag{4-1}$$

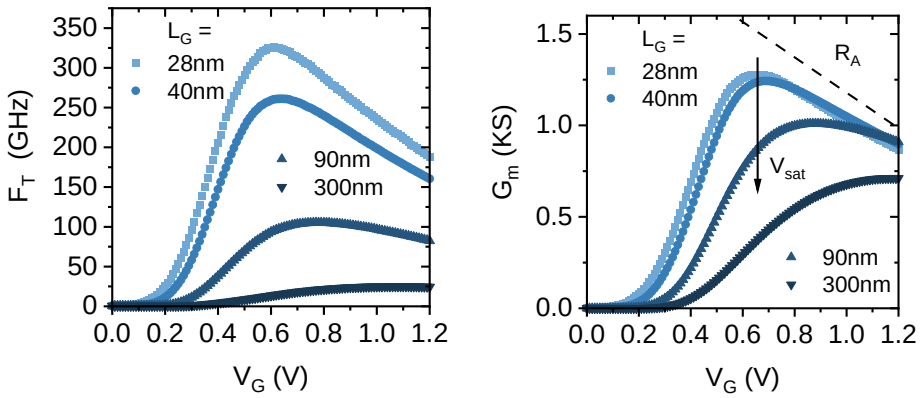
$$f_{max} = \frac{f_T}{2\sqrt{R_{GG}(G_{DS} + 2\pi f_T C_{GD})}} \tag{4-2}$$

From these equations, it is possible to note that f_t is mostly *technology-dependent*. It is mainly related to carrier mobility, access resistance through g_m, and to intrinsic and parasitic device capacitances, through C_{GG}. Therefore, the value of f_t is a good description of the device and can be easily compared between different technologies at same gate length. On the order hand, f_{max} is also *dependent on the design* of the device through the gate resistance R_{GG} for instance [REF]. Therefore, a comparison between technologies is no longer straightforward: the different transistors can have gate width and geometries and gate interconnection designs that change

between foundries. In the following, a deep analysis of f_t and f_{max} frequencies will be done in our LTB FD-SOI technology.

4.2.1 Ft frequency

Erreur ! Source du renvoi introuvable.-right presents the extracted values of the cut-off frequency versus gate bias and channel length at $V_D = 1V$. Similarly to the transconductance gain also shown in Fig., the peak of f_T decreases with smaller channel lengths. This is consistent with eq. 4-1, for which f_T appears rely on the gate transconductance (g_M) and the total gate capacitance (C_{GG}). Thus, to understand the true dependence of f_t with gate length, it is necessary to analyze more in details how both parameters can be first modeled.



4.2.1.1 Transconductance g_m

For frequencies below $(2\pi R_B C_{BB})^{-1}$ and $(2\pi R_G C_{GG})^{-1}$, the drain and source are not capacitively coupled to the gate and substrate. Therefore, g_m can be described by the classical DC current model [XX] recalled in eq. 4-3. The intrinsic transconductance gain on saturation $g_{mi\,sat}$ is only limited by the saturation velocity (V_{sat}). This classic model predicts a maximum value of $g_{mi\,sat}$ that is obtained for short channel devices and strong gate biases. For short channel devices ($L_G < 30\ nm$), $g_{mi\,sat}$ is no longer dependent on V_G and is limited by $W_{eff} C_{ox} V_{sat}$ and can be approximated by eq. 4-4. For non-ideal devices, the source access resistance (R_a) must be included on the model (eq. 4-5). This access resistance component has a greater impact on $g_{mi\,sat}$ for high values of the overdrive voltage, which makes the transconductance gain decrease at strong gate biases.

$$g_{mi\,sat} = \frac{1}{2}\frac{W_{eff}}{L_{eff}} C_{ox}\mu(V_G - V_T)\frac{\frac{\mu(V_G - V_T)}{2V_{sat}L_{eff}} + 2}{\left(\frac{\mu(V_G - V_T)}{2V_{sat}L_{eff}} + 1\right)^2} \quad (4\text{-}3)$$

$$g_{mi\,sat} = W_{eff}C_{ox}V_{sat} \quad for\ L_{eff} \ll \frac{\mu(V_G - V_T)}{4V_{sat}} \quad (4\text{-}4)$$

$$g_{m\,sat} = \frac{g_{mi\,sat}}{\left(g_{mi\,sat}\frac{\mu(V_G - V_T)}{R_a} + 1\right)^2} \quad for\ L_{eff} \ll \frac{\mu(V_G - V_T)}{4V_{sat}} \quad (4\text{-}5)$$

Erreur ! Source du renvoi introuvable. presents the variation of transconductance versus device length in saturation for a High Thermal Budget (HTB) device. By reducing the channel length, the effects of saturation velocity and access resistance become significant, and $g_{m\,sat}$ is no longer linearly dependent on ${L_G}^{-1}$. By fitting the data with this model, we can see the contribution of each parameter to the total current. The carriers scattering on the silicon lattice imposes a limit on the drift velocity of carriers at the channel (V_{sat}~10^7 cm/s) [Taur]. As a result, for short channel lengths, the maximum transconductance becomes theoretically independent from the physical gate length. It is important to note that velocity overshoot occurs in nanoscale transistors as the gate length becomes comparable to the mean free path of the carriers [XX]. Therefore, the value of V_{sat} is mostly treated as a fitting experimental parameter on eq. 4-3, that can be finally bigger than its theoretical limit. In addition, saturation velocity is directly related to the low field mobility [Sun et al., 2007] and thus, can be improved on strained channels where the effective mass of majority carriers is reduced.

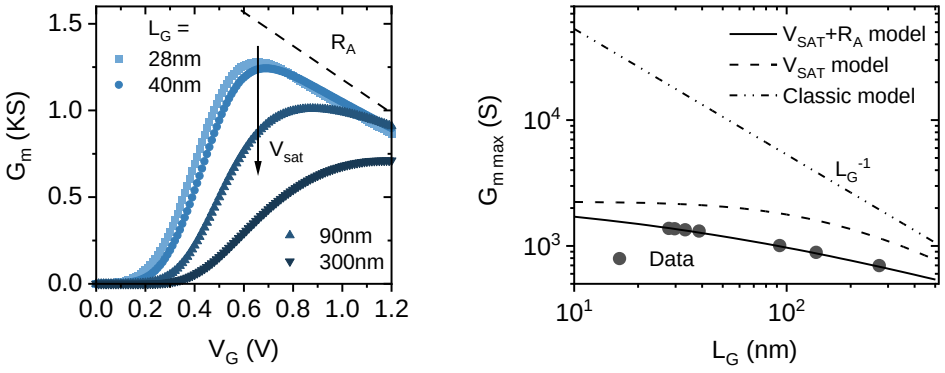


Figure 4-4(left) Measured G_M over V_G in saturation for devices with different gate lengths. Besides the increase on magnitude, the position of the G_M peak shifts toward smaller values of V_G with gate length

reduction. (right) Maximum of transconductance gain (G_M) over device length. The deviation from the $L_G{}^{-1}$ dependence from the classic model is well explained when considering the saturation velocity and access resistance components.

The access resistance of the source terminal has different effects on long and short channel devices (see fig. 4-4). For long devices, it degrades $g_{mi\ sat}$ at high values of V_G, shifting the position of the transconductance maximum closer to the threshold voltage. In addition, the access resistance decreases the maximum of transconductance in respect to the saturation velocity limit. Because of those effects, the gain on saturation from 30 nm gate length devices is near one decade smaller than what would be expected from the $L_G{}^{-1}$ trend.

4.2.1.2 Gate capacitance

As discussed in the previous chapters, the total gate capacitance C_{GG} is the sum of the gate oxide capacitance, the inner and outer fringe capacitances and of external parasitic components. From two-ports high frequency measurements, it is possible to separately extract the capacitance between the gate and the source (C_{GS}) and the drain (C_{GD}) as presented in **Erreur ! Source du renvoi introuvable.**-right. At low V_{DS}, both are symmetrical but, in the saturation regime, due to the channel pitch-off, C_{GD} becomes smaller than C_{GS} as V_{DS} increases.

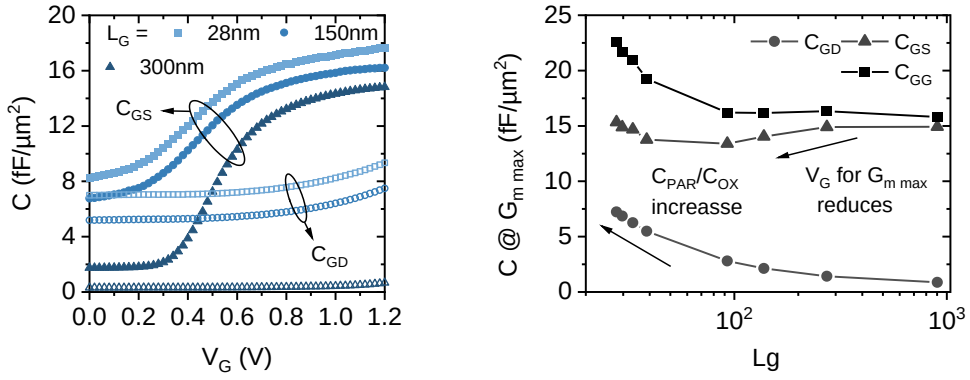


Figure 4-5(left) Measured C_{GS} and C_{GD} over V_G in saturation for devices with different gate length. While C_{GS} is due to the inversion and outer capacitance components XXX, C_{GD} is mostly given by the outer and inner capacitance components. (right) Gate to Drain capacitance (C_{GD}), Gate to Source capacitance (C_{GS}) and total gate capacitance (C_{GG}) extracted at the maximum of G_M as the function of gate length. The values of capacitance are normalized by the gate surface

Erreur ! Source du renvoi introuvable.-left presents each component normalized by the gate surface and extracted at the maximum of the transconductance. C_{GS} is the sum of the pinched inversion charge capacitance and the outer fringe capacitance with the source. For long devices, the parasitic components are negligible

and C_{GS}/W remains constant at same overdrive bias. On the other hand, as the channel near the drain is depleted in saturation, C_{GD} is only consists of the inner and outer fringe capacitances. Those parasitic components become significant as the channel length decreases and are responsible to near half of C_{GG} at the maximum of G_m when L_G reaches 30nm.

4.2.1.3 max ft vs L_G

Erreur ! Source du renvoi introuvable.-left reports the maximum of f_T as function of the gate length. The factor α allows to characterize the dependence of $F_T$on L_G By dividing F_T by the device length, it becomes clear that the value of α is different between short and long channels. For long ideal transistors, α is equal to 2 as G_M and ${C_{GG}}^{-1}$ linearly increase with ${L_G}^{-1}$. However, for shorter devices, the degradation of G_M and relative increase of parasitic resistances and capacitances decrease the value of α until $\alpha < 1$ for $L_G < 40\,nm$. While α is greater than zero, a reduced channel length is still advantageous to improve this FoM

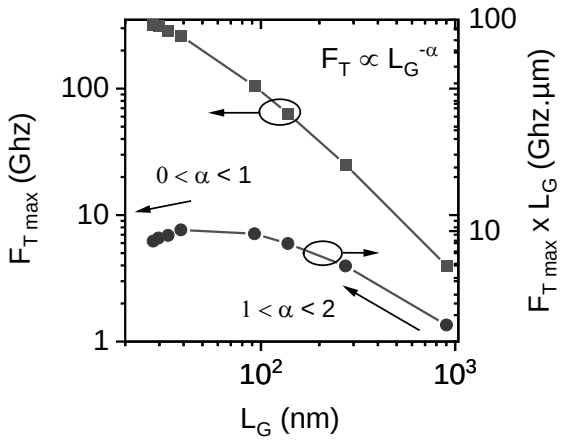


Figure 4-6 Measured maximum of the cut-off frequency (F_T) vs the gate length. For long L_G, ….Fshort L_g

4.2.2 F_{max} frequency

From eq. 4-2, f_{max} is dependent on the same components as f_t, in addition to the channel conductance G_{DS} and the gate resistance R_{GG}.

4.2.2.1 Channel conductance G_{ds}

The effects of channel length reduction on the channel conductance G_{DS} is presented in **Erreur ! Source du renvoi introuvable.**. As for G_M, it increases with channel length reduction, but it is also affected by the saturation velocity and access

resistances. For a fixed gate bias, it follows a power law dependence with gate length with an exponent smaller than one. However, as G_{DS} decreases on saturation regime, the term $2\pi f_t C_{gd}$ has a greater impact on f_{max}. This effect can be observed in **Erreur ! Source du renvoi introuvable.**-right where G_{DS} is responsible for only one fifth of the total contribution on f_{max}.

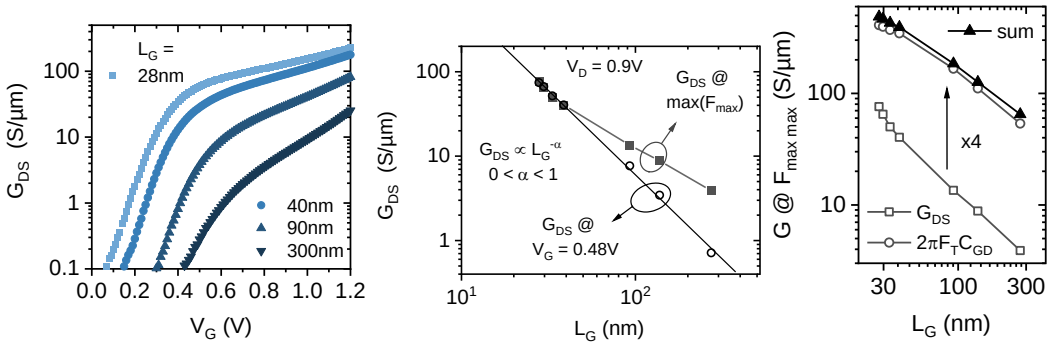


Figure 4-7 (left) Measured G_{DS} over V_G in saturation for devices with different gate lengths (middle) Channel conductance (G_{DS}) versus gate length at fixed V_G and at the maximum of F_{MAX} (right) Impact of G_{DS} and C_{GD} on max oscillation frequency (F_{MAX}) equation as function of gate length. The channel conductance has very a small contribution to the degradation of F_{MAX}.

4.2.2.2 Gate resistance R_{GG}

The gate resistance is given by the sum of three components: a fixed contact resistance (R_{CT}), a horizontal (R_H) and a vertical resistance (R_V) that present different dependences on the device dimensions and number of gate contacts (eq. 4-6). **Erreur ! Source du renvoi introuvable.**-middle presents the dependence of gate resistance with channel length for a device with 1 µm gate width and a single gate contact. In accordance to eq. 4-6, the gate resistance increases with the reduction of the channel length. Differently from what is observed on f_t, the optimal value of f_{max} is not obtained for the smallest gate length.

For the case of this device, the dependence on ${R_{GG}}^{1/2}$ actually decreases f_{max} for $L_G < 40\ nm$. Finally, the model for f_{max} can be tested against measurements of each parameter. **Erreur ! Source du renvoi introuvable.**-right presents the dependence of f_{max} with gate length. The model is calculated using eq. 4-2 and the previously extracted parameters at given gate bias. It presents a good agreement with the data extracted from the U_{21} gain.

$$R_{gg} = \begin{cases} R_{ct} + \dfrac{R_H}{3}\dfrac{W}{Lg} + \dfrac{R_V}{Lg \cdot W}, & 1\ contact \\ \dfrac{R_{ct}}{2} + \dfrac{R_H}{12}\dfrac{W}{Lg} + \dfrac{R_V}{2 \cdot Lg \cdot W}, & 2\ contacts \end{cases} \quad (4\text{-}6)$$

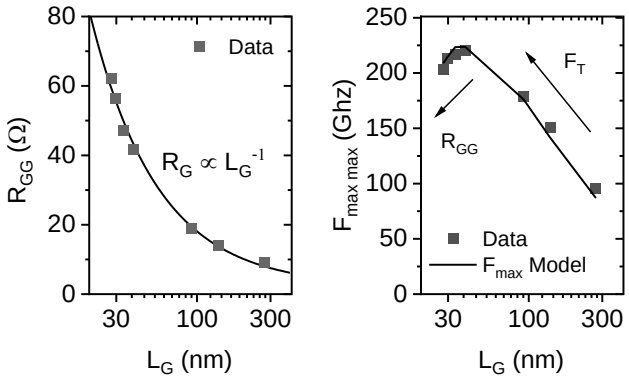


Figure 4-8 (left) Gate resistance (R_{GG}) for a single finger as function of gate length. (right) Model and measurement of F_{MAX} over gate length.

4.3 SHORT CHANNEL MOBILITY

4.3.1 Mobility extraction with a revisited CV-split method

Traditionally, channel mobility is extracted from long devices using capacitance and current measurements with the CV-split method. The capacitance is used to extract the density of inversion charges, while the current is used to obtain the channel resistivity. As a reminder, if we consider an ideal MOSFET device without access resistance and with the source connected to the ground, the drain current in the linear regime can be expressed by eq. 4-7. This model is based on the only assumption that the drain bias is sufficiently small so that the density of free carriers is constant over the channel length:

$$Id_{lin} = \frac{W}{L_G}\mu_{eff} \cdot \int_0^{L_G} Q_y \cdot \overrightarrow{E_y} \cdot dl \cong \frac{W}{L_G}\mu_{eff} \cdot Q_{inv} \cdot Vd_{lin} \qquad (4\text{-}7)$$

The channel capacitance C_{GC} must be measured between the gate and the source and the drain while keeping the substrate floating as depicted in FIG. Its expression is then given by eq. X where C_{ox} and $C_{inv} = dQ_{inv}/d\Psi_s$ are the gate oxide and the carrier inversion capacitance normalized by the device surface respectively; and where C_{if} denotes the inner fringe capacitance normalized this time by the device width. For a sufficiently long channel, where $C_{if} + C_{par} \ll C_{inv}$ for $V_G > V_0$, C_{ch} can be only expressed on C_{inv}. The inversion charge is then extracted from the integral of the gate capacitance as given by eq. 4-9. Then combining eq. 4-9 with eq. 4-10, we obtain the value of the mobility as a function of the gate bias.

$$C_{ch} = \frac{WL_G}{{C_{ox}}^{-1} + \left(C_{inv} + \frac{C_{if}}{L_G}\right)^{-1}} + C_{par} \cong \frac{WL_G}{{C_{ox}}^{-1} + {C_{inv}}^{-1}} \qquad (4\text{-}8)$$

$$Q_{inv} = \frac{1}{WL_G}\int_{V_0}^{V_G} C_{ch} \cdot dV \quad where: V_0 \ll V_T \qquad (4\text{-}9)$$

$$\mu_{eff} = {L_G}^2 \cdot \frac{Id_{lin}}{Q_{inv} \cdot Vd_{lin}} \qquad (4\text{-}10)$$

This method is useful to extract the intrinsic mobility of long devices and to identify the physical mechanisms responsible for mobility degradation, that are not dependent on the channel length. For devices with thin gate oxides, it is also important

to measure the current for both positive and negative values of Vd_{lin}, in order to correct the raw data from gate leakage current and V_{FB} shift artifacts.

For short channel devices, the access resistance and parasitic capacitances are no longer negligible; and this method fails to provide an accurate description of the intrinsic mobility. To overcome this issue, the CV-split method was improved to allow mobility extraction on short channel devices. A previous improved CV-split method was presented by [K. Romanjek], where it is proposed to use C_{GC} measurements from devices with different gate lengths to estimate the inversion charge of each device. However, the proposed method is limited for 3 reasons (1) it does not account for the variation of inner fringe fields with gate length (2) it considers the charge on the channel to be constant along the whole structure. (3) the method proposed to extract the real gate length from the devices is not feasible on FD-SOI structures.

The revisited method proposed in this work aims to improve the accuracy of the mobility extraction, based on a new de-embedding able to completely remove the parasitics affecting the capacitance and current measurements. The new technique is based on the use of two devices with different channel lengths to extract the mobility at the center of device channel.

The first step is the extraction of Q_{Inv}. To do that, we assume that

- the density of dopants is constant over channel length so that the channel carrier density at any bias is the same whatever the channel length as illustrated in **Erreur ! Source du renvoi introuvable.**,
- the junction profile is identical whatever the device size so that the same access resistance and capacitance is obtained in short and long devices when $V_{DS} = 0V$.

Such assumptions are generally easily verified in FD-SOI symmetrical MOSFETs.

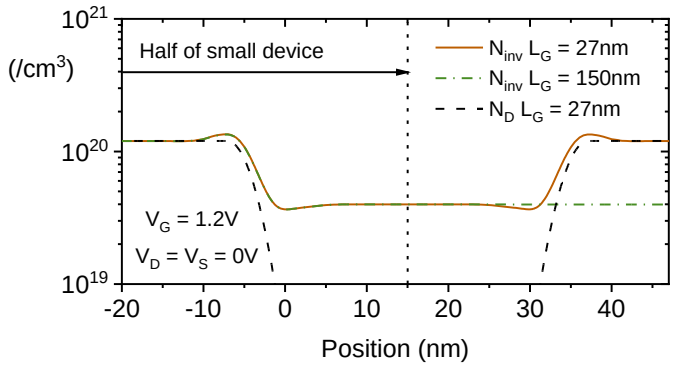


Figure 4-9 Dopants and electron charge distribution on two NMOS devices with different channel lengths.

Based on these hypotheses, the inversion capacitance of long device is easily extracted by subtracting the access capacitance assessed on the shorter device. Basically, we calculate the differential capacitance (C_{diff}) between two devices with channel lengths L_{G1} and L_{G2} as shown by eq. 4-11. In the short channel length case, the parasitic component must be considered, and C_{diff} is equivalent to the gate capacitance of a device with channel length $L_1 - L_2$ with $L_1 > L_2$, while C_K (eq. 4-12) is the remaining inner fringe capacitance that cannot be completely removed from this simple subtraction.

$$C_{diff} = C_{ch1} - C_{ch2} = \frac{W(L_1 - L_2)}{\frac{1}{C_{INV} + C_K} + \frac{1}{C_{OX}}} \tag{4-11}$$

$$Where: C_K = \frac{L_1 L_2 (C_{INV} + C_{OX})\left(C_{if1} - C_{if2}\right) + C_{if1} C_{if2} (L_1 - L_2)}{L_1 L_2 (C_{INV} + C_{OX})(L_1 - L_2) + C_{if1} {L_1}^2 - C_{if2} {L_2}^2} \tag{4-12}$$

For the majority of realistic FD-SOI MOSFET with a reasonable amount of parasitic, the inner fringe capacitance of each device is much smaller than the oxide capacitance. On those cases, the expression of C_K can be approximated by eq. 4-13. If $C_{if1} = C_{if2}$ is considered to be constant over gate length, C_K is equal to zero. However, it is known that the fringe capacitance increases with channel length.

$$C_K = \frac{C_{if1} - C_{if2}}{L_1 - L_2} \quad for \quad C_{OX} \gg C_{ifn} \tag{4-13}$$

Finally, we can remove C_K from C_{diff} and obtain the curve of Q_{inv} using eq. 4-14 and eq. 4-15. V_{HIGH} is the gate bias that gives a value of capacitance in strong inversion regime, when $C_{INV}(V_{HIGH}) \gg C_{ox}$. This value is obtained for the point in

C_{diff} where the differential capacitance curve flattens, and is equivalent to the oxide capacitance C_{ox}. Eq. 4-14 considers that $C_{if1} - C_{if2}$ is constant over V_G. This approximation is valid even if C_{if1} and C_{if2} vary with de gate bias. Despite the different channel lengths, both devices have identical junction profiles. Therefore, the difference between the inner fringe capacitances at any gate bias is going to be approximately the same, independent on the depletion of the access junction. This can be verified from the C_{diff} curve, as differently from the channel capacitance, it is supposed to be flat on depletion regime.

$$C_{INV\ diff} = \frac{1}{\frac{1}{C_{diff}} - \frac{1}{C_{diff}(V_{HIGH})}} - \frac{1}{\frac{1}{C_{diff}(V_0)} - \frac{1}{C_{diff}(V_{HIGH})}} \tag{4-14}$$

$$C_{CH\ diff} = \frac{1}{{C_{INV}}^{-1} + C_{diff}(V_{INV})^{-1}} \tag{4-15}$$

$$Q_{ch\ diff} = \int_{V_0}^{V_G} C_{CH\ diff} \cdot dV \tag{4-16}$$

This procedure is also useful for the extraction of the inversion charge from devices where the substrate is physically connected to the source. In this case, the box capacitance cannot be avoided and C_K is given by eq. 4-17 The method is still valid to extract the box capacitance from a single long device and the charge can be obtained from the same equations as before where C_{diff} can be replaced directly by the gate capacitance.

$$C_K = \frac{C_{if1} - C_{if2}}{L_1 - L_2} + C_{box} \tag{4-17}$$

The second step is to remove the parasitic resistance from the current measurements. For that, we consider both devices, at given gate bias, as a series resistance network sharing the same elements for the access region. The equivalent current for the ideal device with channel length $L_1 - L_2$ is given by the difference of channel conductance between the both devices as presented in equation 4-17.

$$Id_{lin\ diff} = \frac{1}{{Id_{lin1}}^{-1} - {Id_{lin2}}^{-1}} \tag{4-18}$$

With access resistance and parasitic capacitances extracted from both devices, the mobility is obtained similarly to the classic method with eq. 4-10. The only difference is that L_G must be replaced by $L_1 - L_2$ for this case.

4.3.2 Mobility extraction from high frequency measurements

The same method can be done using high frequency measurements performed by a VNA. In that case, the capacitance and conductance of the device are extracted from a single S-parameters measurements. The gate capacitance (C_{GG}) is extracted from the imaginary part of Y_{11} , while the channel conductance (G_{DS}) is extracted from the real part of Y_{22}, as illustrated in **Erreur ! Source du renvoi introuvable.**. The device is measured at $V_D = V_S = 0$ for different values of gate bias.

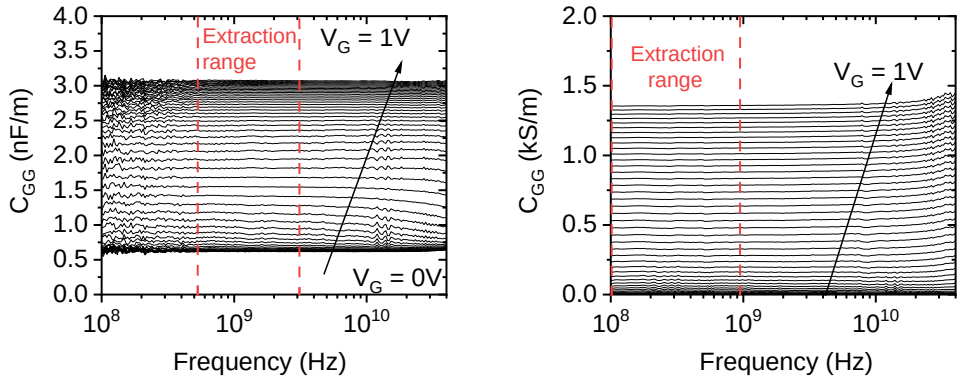


Figure 4-10 Capacitance and conductance over frequency at different gate biases obtained from S-parameters measurements. The dashed lines indicate the frequency range used on the extraction of each parameter. (Left) Gate capacitance (C_{GG}). (Right) Channel conductance (G_{DS})

High frequency characterizations provide some benefits in comparison to the classic DC-LF coupled measurements for the application of this method. Indeed S-parameters measurements allow measuring the device with zero bias in the drain. Without any DC current on the channel, the density of carriers over the channel is perfectly homogenous and symmetrical. Moreover, both capacitance and conductance are extracted at the same time and with the same bias, which eliminates any difference in the threshold voltage between measurements. In addition, the extraction of capacitance from high frequency VNA measurements are much more accurate than low frequency measurements performed with a classic capacimeter. At 1 GHz, it is possible to obtain low noise capacitance curves from devices with a gate surface as small as 5 μm^2. Finally, high frequency measurements are mostly immune to self-heating effects, which can be useful for the extraction of mobility at cryogenic temperatures.

After obtaining the capacitance and conductance curves, the steps are the same as the previous method. The only difference is that the conductance $G_{ds} \approx Id_{lin}/Vd_{lin}$ is obtained directly from the measurements.The mobility, in that case, is obtained using eq. 4-19.

$$\mu_{eff} = (L_1 - L_2)^2 \cdot \frac{G_{DS\,diff}}{Q_{diff}} \tag{4-19}$$

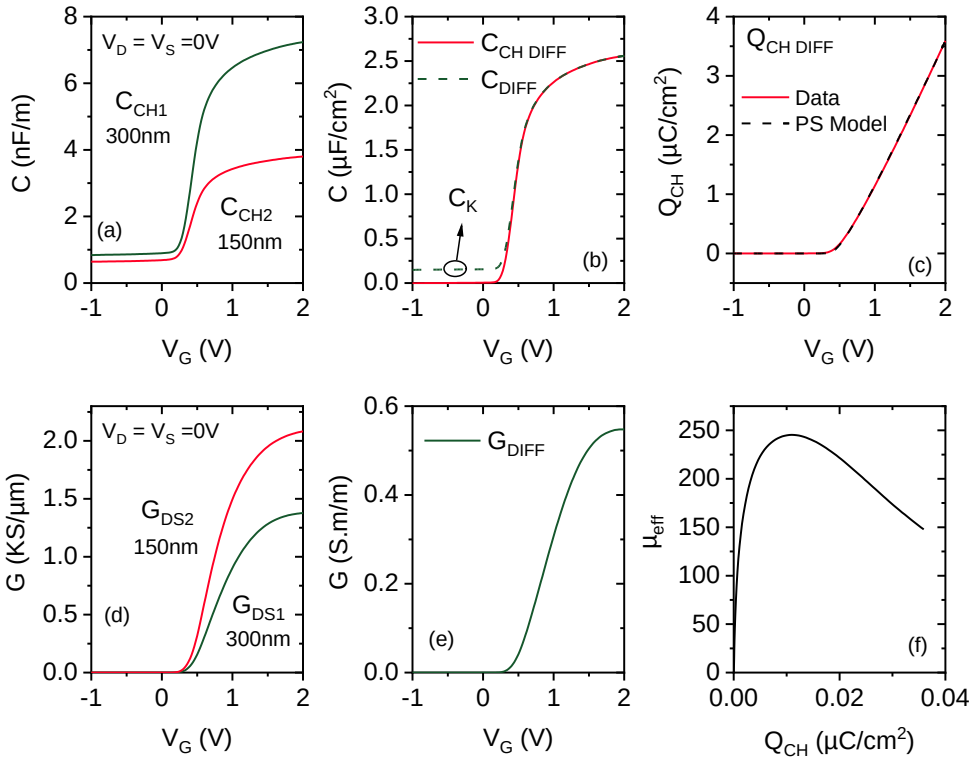


Figure 4-11 Example of extraction of channel mobility using s-parameter measurement of two devices with gate lengths $L_1 = 300$nm and $L_2 = 150nm$. (a) Gate to channel capacitance from devices 1 and 2. (b) Differential capacitance and channel capacitance of an equivalent device with gate length $L_{diff} = L_1 - L_2$. (c) Inversion charge from the differential capacitance compared to a 1-D Poisson-Schrodinger model of the device. (d) Channel conductance from devices 1 and 2. (e) Differential conductance. (f) Effective mobility from the equivalent L1–L2 device.

Erreur ! Source du renvoi introuvable. presents the different steps necessary for the application of the revisited CV-split method. From the capacitance curves (C_{CH1} & C_{CH2}) of two devices with different lengths, we extract the differential capacitance (C_{diff}). Then, the remaining fringe capacitance (C_K) on depletion is removed from the C_{diff} using eq. 4-14 and eq. 4-15 to obtain $C_{CH\,diff}$. This parameter is equivalent to the channel capacitance of a device with channel length $L_1 - L_2$. The charge ($Q_{CH\,diff}$) is then obtained from the integral of $C_{CH\,diff}$ over gate bias. In the example in **Erreur ! Source du renvoi introuvable.**, the charge is compared to a Poisson-Schrodinger 1-

D capacitive model of a device with same Equivalent Oxide Thickness (EOT). The similar results demonstrate that it is possible to obtain the charge characteristics from the device without parasitic short channel effects using the revisited method. From the conductance curves (G_{DS1} & G_{DS2}), we obtain the differential conductance G_{diff} using eq. 4-18. Finally, the mobility is extracted from $Q_{CH\ diff}$ and G_{diff} using eq. 4-19.

After the extraction of the mobility curve, we can fit it using the universal model for mobility degradation presented in equation. 4-20 to equation. 4-23 [Takagi]. Three mobility components are function of the charge density in this model: The coulomb scattering component (μ_{coul}) is related to near interface charges of the gate oxide and dominates at low charge density. Phonon scattering (μ_{phonon}) is explained by the collision with the silicon lattice and dopants, and is constant as a function of the charge. Finally, interface roughness scattering (μ_{rough}) prevails at high Q_{inv} where a high density of carriers are pushed towards the oxide interface.

$$\frac{1}{\mu_{eff}} = \frac{1}{\mu_{coul}} + \frac{1}{\mu_{phonon}} + \frac{1}{\mu_{rough}} \tag{4-20}$$

$$\mu_{coul} = 10^{\log(N_{inv})A_c + B_c} \tag{4-21}$$

$$\mu_{phonon} = 10^{\log(N_{inv})A_p + B_p} \tag{4-22}$$

$$\mu_{rough} = 10^{-\log(N_{inv})A_r + B_r} \tag{4-23}$$

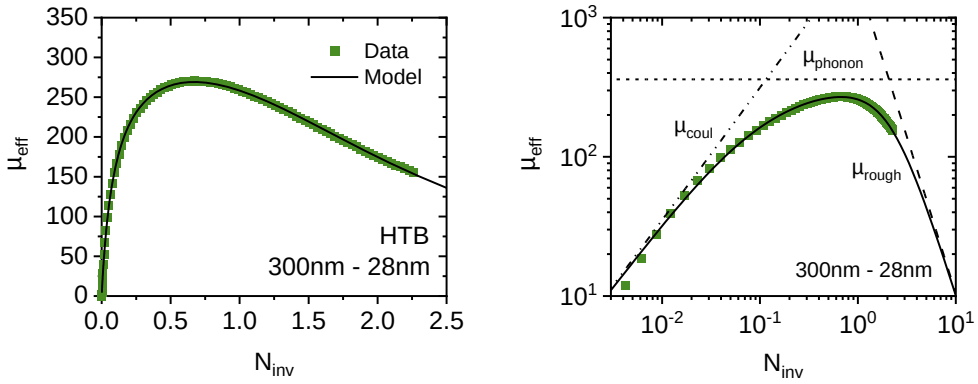


Figure 4-12 Mobility extracted from devices with channel length of 300nm and 28nm compared to the fitted model. (left) Data plotted on linear-linear (right) Data plotted on log-log

The mobility characteristics extracted from 300-nm and 28-nm devices is well fitted by the model, as illustrated in **Erreur ! Source du renvoi introuvable.**. In particular, the coulomb mobility component is fitted with an exponent A_C of one. Such a valueees with the underlying physics scattering mechanism. [ref: mike told me]. This

demonstrates that the improved accuracy of the revisited CV-split method can be useful to explore the effects of mobility at very low charge densities and with different channel lengths.

It is possible to apply this method to different couples of gate lengths and understand the effects of the variation of mobility along the channel. **Erreur ! Source du renvoi introuvable.**-left presents the mobility curves obtained from this method when L1 changes from 300 nm to 100 nm. Clearly, the mobility increases as L1 decreases. A similar effect is observed by changing L2 from 28 nm up to 150 nm, as shown in **Erreur ! Source du renvoi introuvable.**-middle. The smaller is the difference in length, the stronger is the mobility. In addition, the effect is stronger on the phonon scattering region, which is largely affected by strain effects. Indeed, this effect corroborates to the hypothesis of contact etch-stop-layer CESL strain pockets. [L. Pham-Nguyen] suggested that the effect of vertical compressive and tensile strain on the channel created from the Contact Etch Stop Layer (CESL) deposition is stronger on the corners of the gate. Therefore, the strain has a stronger impact on shorter devices.

$$\frac{L_1 + L_2}{\mu_{tot}} = \frac{L_1}{\mu_1} + \frac{L_2}{\mu_2} \tag{4-24}$$

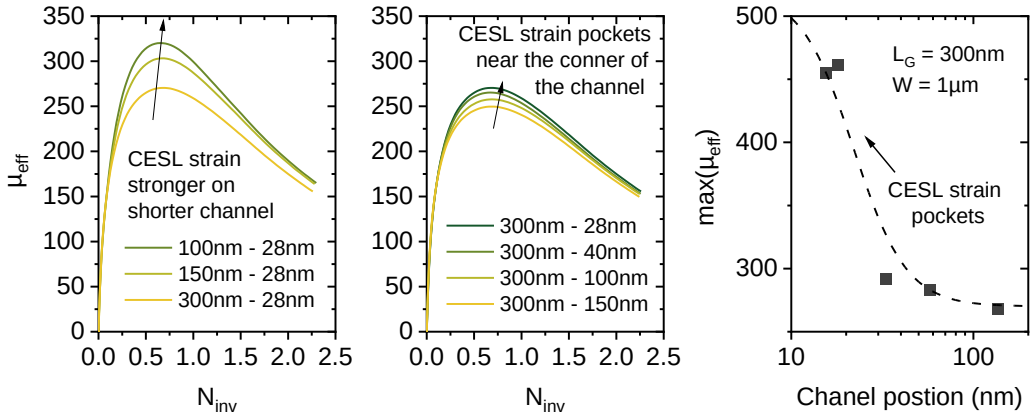


Figure 4-13 Extraction of mobility using different gate length pairs. (left) Variation of L1 while L2 remains constant. (middle) Variation of L2 while L1 remains constant. (right) Estimation of the variation of mobility along the channel from a device with gate length of 300nm plotted along the mobility variation model proposed by [L. Pham-Nguyen].

Finally, considering the channel to be the series combination of chunks of MOSFET devices with same electrostatic characteristics but different motilities, the effective mobility from one portion of the device can be given as a combination of two

portions using the simple equation 4-24. Therefore, it is possible to estimate the mobility from each portion of the channel using the results presented in **Erreur ! Source du renvoi introuvable.**-middle. As expected, a strong mobility pocked is shown in the first 20 nm portion of the channel near the access where the strain is expected to me most effective.

4.3.3 Mobility effect of strained LTB device

For low thermal budget devices, the technique was used with two devices of gate length equals to120 nm and 60 nm, respectively. The effect of the interface anneal and vertical strain on NMOS LTB devices is verified in three different process variants. A comparative study is performed on the mobility effect of hydrogen forming gas (FG) final anneal performed on an old batch and the high-pressure deuterium (HPD_2) anneal used for the most recent process flow. In addition a third variant is fabricated with a vertical direction, tensile strain from the SIN contact etch-stop-layer (CESL) that was highly doped with hydrogen, similarly to [L. Pham-Nguyen].

A. V1: FGA + Neutral CESL strain
B. V2: HPD_2 + Neutral CESL strain
C. V3: HPD_2 + Tensile CESL strain

Figure X. present the mobility for each variant extracted using the revisited C-V split method. On the left, the mobility curve obtained from variant V2 is presented superposed to the fit from eq. X with each of its components. In **Erreur ! Source du renvoi introuvable.**–middle, mobilities from variants V1 and V2 are compared. Between HPD_2 and FGA anneals, there is a small reduction on the density of gate oxide interface states, from $8 \cdot 10^{11}\ cm^{-2}$ to $3 \cdot 10^{11}\ cm^{-2}$. As expected, the interface improvement results on the improvement of mobility at low density of inversion charge from the reduction of coulomb scattering. This effect is not apparent in strong inversion regime where other scattering mechanisms become predominant. Despite this, the increase in mobility occurs near the region where the transconductance gain is most significant.

Strain techniques that introduce compressive or tensile constrictions on the silicon lattice in the transistor channel provide an elegant method of improving carrier's mobility. From the band splitting and carrier effective mass change, the impact of strain on the carrier mobility from phonon scattering reduction is well

documented in the literature [refs+Min Chu]. **Erreur ! Source du renvoi introuvable.**-right presents the effect of the mobility boost from the tensile strain. Between variants V2 and V3, there is a noticeable increase of mobility at medium and high charge densities.

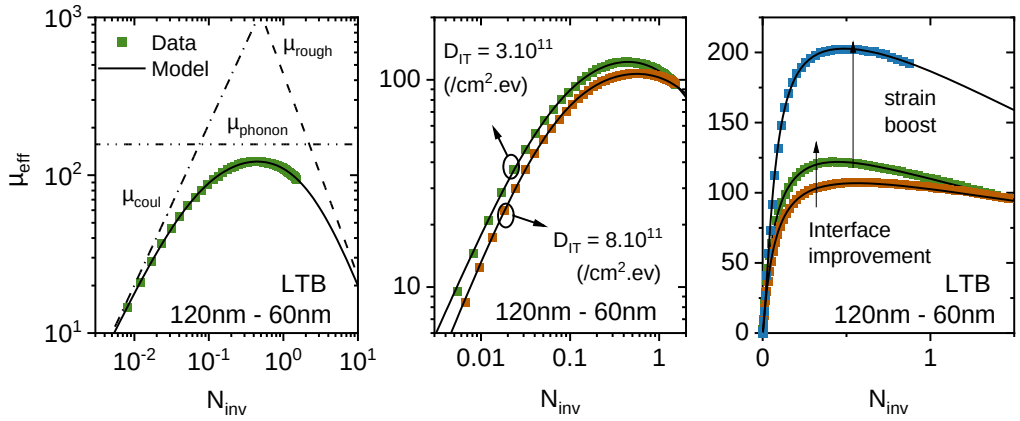


Figure 4-14 (left) Mobility extracted from LTB device that was exposed to an HPD_2 anneal using revisited method. (middle) Dependence of mobility at low density of charges as function of interface state density obtained from devices exposed to an HPD_2 and a classic FG anneal. (right) Effect of boost and interface improvement on mobility of LTB devices.

The effects of this uniaxial strain on the high frequency performance of the device are clear. **Erreur ! Source du renvoi introuvable.** shows a comparison of the effects of a mobility improvement on the device's FoMs. While G_M increases for any drain and gate bias, G_{DS} is impacted only in linear regime. This behavior is very helpful as the increase of G_M is directly related to a boost on F_T while the non-variation of G_{DS} does not negatively affect the value of F_{MAX}. **Erreur ! Source du renvoi introuvable.**-right presents the curves of F_T versus drain current for different drain bias on a LTB device with 45-nm channel length. The increase in mobility from the CESL strain has a direct impact on the cut-off frequency. **Erreur ! Source du renvoi introuvable.** provides an explicit picture of this effect where the maximum of F_T and F_{MAX} are plotted against the peak of G_M for a 45-nm device measured from different dies. By increasing G_M, both F_T and F_{MAX} are directly improved.

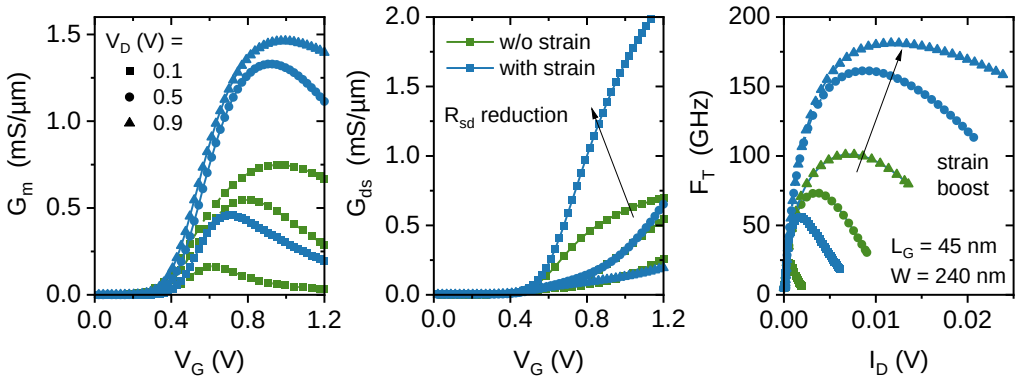


Figure 4-15 Effect of strained channel of RF FoMs for different V_G and V_D bias. (left) Transconductance gain (G_M). (middle) Channel conductance (G_{DS}). (right) Cut-off frequency (F_T).

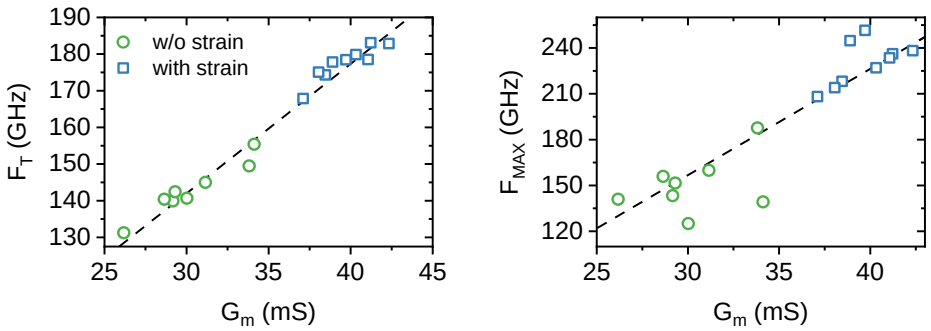


Figure 4-16 Correlation of F_T and F_{MAX} with the boost of transconductance (G_M) for devices with and without a stained channel.

The strain also has a peculiar effect on the devices reliability. Several sources have already demonstrated the benefits of tensile CESL strain on NMOS reliability performance. The strain reduces the band gap and at the same time the minimum conduction band in the inversion layer. As a result, the barrier height for emission of carriers across the gate oxide interface is expected to increase. This effect reduces the probability of electrons tunneling from the inversion layer through the oxide and can be seen as a reduction of the gate leakage current at same overdrive bias [C. Claeys]. On the other hand, the smaller bandgap and lower V_{DSAT} at same gate overdrive bias results in a higher multiplication factor and enhanced creation of electron-hole pairs in saturation regime. Finally, [Hui Ling Huang] also reports an improvement on the quality of the gate stack and substantial reduction of interface states density for devices capped with a tensile CESL. It suggests that the excess of hydrogen on the CESL cap diffuses through the oxide and helps to passivate the oxide interface.

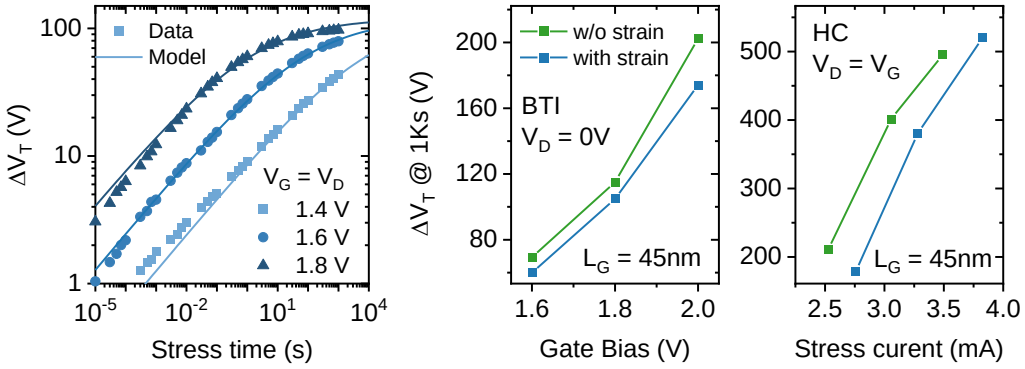


Figure 4-17 Effects of stain on devices reliability. (left) V_T degradation as function of time for different HC stress conditions fitted with a saturated power law model. (right) V_T degradation after 1Ks for both strain conditions as function of stress bias.

To explore the effects of the tensile CESL on the devices reliability FoMs, BTI and HC degradation measurements where performed on NMOS devices and are presented in **Erreur ! Source du renvoi introuvable.**. As example, **Erreur ! Source du renvoi introuvable.**-left presents the drift of saturation V_T from HC degradation performed at different V_{DD} biases on the strained device. The results of each stress mode is presented at the right and are in agreement with similar tests performed in the literature [Hui Ling Huang, C. Claeys]. It is clear that the strained device has improved performance on both BTI and HC degradations. For BTI, the difference on the drift of V_T at same stress condition can be explained by the reduced tunneling probability from the lowered conduction band. For HC degradation, the V_T drift is actually unaltered between devices at same bias condition. However, because the stress current is improved from the strain, the CESL provides a benefit on HC degradation at constant drain current. This invariability is not completely understood. The generation of hot carriers is expected to increase from higher creation of electron-hole pairs, which in consequence should negatively affect the HC degradation. However, the smaller carrier's effective mass and increased barrier height could compensate to hinder the oxide degradation.

4.4 LOW PERMITIVITY SPACERS PERFORMANCE

As discussed previously, the contribution of parasitic capacitance between gate and source/drain terminals increases with gate length scaling. As a consequence, integrating the spacers using an insulator material with low permittivity (low-k) is an smart method to improve not only the RF FoMs of the transistor, but also to reduce the load capacitance of standard cell for digital applicaton. Silicon oxide (SiO_2) has arelative dielectric constant (k or ε_r) of 3.9 whereas silicon nitride (SiN) has an relative permittivity of approximately 7. Despite the higher dielectric constant, silicon nitride still the most common material for spacer applications because it provides higher etching selectivity than the one of the thick dielectric used for the contact. On the other hand, so far low-k materials are only integrated in the back-end of line to reduce interconnection parasitic. Nevertheless, the use of these materials for the gate spacers recently gains a lot of attention for future advanced CMOS nodes and high performance RF devices. **Erreur ! Source du renvoi introuvable.**-right shows the permittivity of the materials, that are expected to be used for the gate isolation on advanced nodes from the 2022 International Roadmap for Devices and Systems.

Preferably, the spacer should be a material, that can be etched selectively to the gate dielectric layer, and that does not interfere with the epitaxial growth of source and drain. SiCO is a great candidate for low temperature applications despite its trapping characteristics. **Erreur ! Source du renvoi introuvable.**-left presents the benefit in the outer fringe capacitance as function of gate length brought by SiCO compared to SiN. Because the fringe component does not scale with gate length, the spacer contribution increases for advanced nodes.

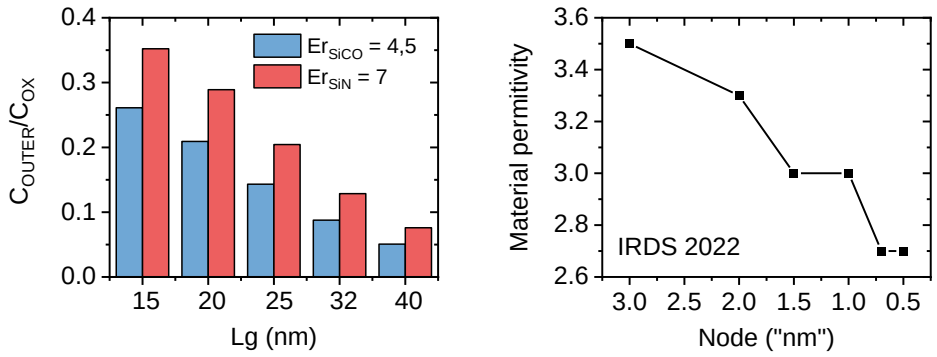


Figure 4-18(left) TCAD simulation showing the contribution of the outer capacitance to the gate oxide capacitance (right) 2022 International Roadmap for Devices and Systems for the spacer permittivity of advanced CMOS nodes.

To quantify the improvement on the parasitic capacitance, a comparative study is performed between HTB devices and LTB devices integrating SiN and SiCO spacer respectively. **Erreur ! Source du renvoi introuvable.** shows that, switching from SiN to SiCO improves the off state ($V_G = 0V$) capacitance by a factor between 25% and 30%. Note that the different values of parasitic capacitance between PMOS and NMOS devices arises from the different thicknesses of spacers and the variation in junction position.

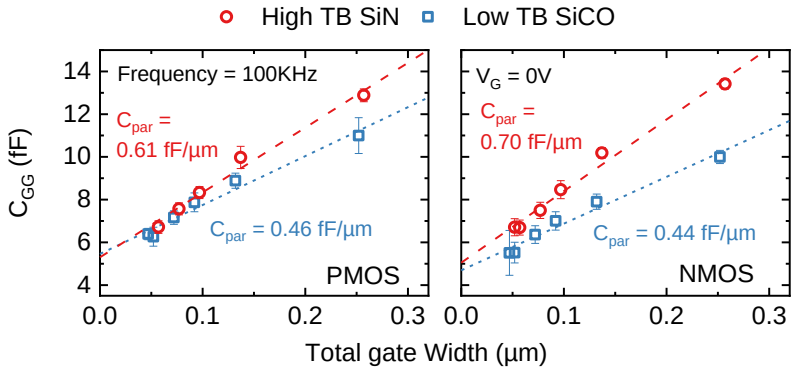


Figure 4-19 Low frequency gate capacitance in off-state ($V_G = 0$) plotted against gate width. $C_{par}=C_{if}+C_{of}+C_{2D}{}^{w}+C_{3D}$ From the variation over width, it is possible to extract the parasitic capacitance of the 2D width dependent device and exclude the 3D fixed component.

In addition, the parasitic capacitance of HTB and LTB NMOSFET as a function of gate voltage is extracted in **Erreur ! Source du renvoi introuvable.**. As a reminder, the extraction technique consists, firstly in calculating the differential capacitance technique between two Lg in order to obtain the channel capacitance $C_{GC}(V_G)$ [ref]; and then in subtracting the calculated C_{GC} from the total gate capacitance C_{GC} to deduce C_{par}. In **Erreur ! Source du renvoi introuvable.**-left, it is clear that the SiCO spacer device exhibits much lower fringe capacitance whether the gate bias. For negative values of overdrive bias, this parasitic capacitance consists of two main components: C_{outer} and C_{inner}. As the gate bias increases $V_G>V_T$, the inner component cancels because of the formation of the inversion layer, and the parasitic capacitance is therefore mainly given by its outer component. For high overdrive bias, the gain on parasitic is near 30%. Considering that in strong inversion regime most of the outer capacitance is associated to the spacer capacitance, we evaluate that the permittivity of SiCO is 35% lower than the one of SiN.

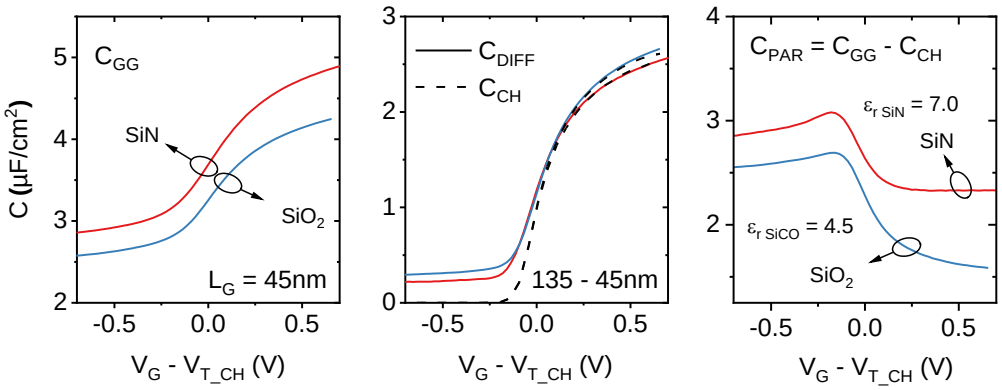


Figure 4-20 Extraction of parasitic capacitance as function of V_G from two devices with different materials for the spacer oxide (SiN and SiCO). (Left) Gate capacitance (C_{GG}) obtained from RF measurements of 45nm devices with each spacer material variant. (middle) Differential and channel capacitance obtained from devices with $L_{G1} = 135nm$ and $L_G = 45nm$. (right) Parasitic fringe capacitance estimated for each variant.

Those benefits are even better in saturation regime. **Erreur ! Source du renvoi introuvable.** illustrates the gate capacitance and its components as the function of V_G and V_D for both HTB and LTB devices. Figure X-left presents the C_{GS} versus overdrive voltage. On this case there is not a big difference at V_D=1V between both devices since C_{GS} is dominated by the inversion charge capacitance C_{GC}~C_{inv}, which is similar for HTB and LBT transistors. However, for C_{GD}, the difference between both devices is much larger, as shown in **Erreur ! Source du renvoi introuvable.**-middle. Because the drain is depleted at V_D=1V, C_{GD} is mainly ascribed to the outer and inner fringe capacitances. This results in a clear improvement on the C_{GG} capacitance, highlighted in **Erreur ! Source du renvoi introuvable.**-right, for the SiCO LTB device with respect to the SiN counterpart.

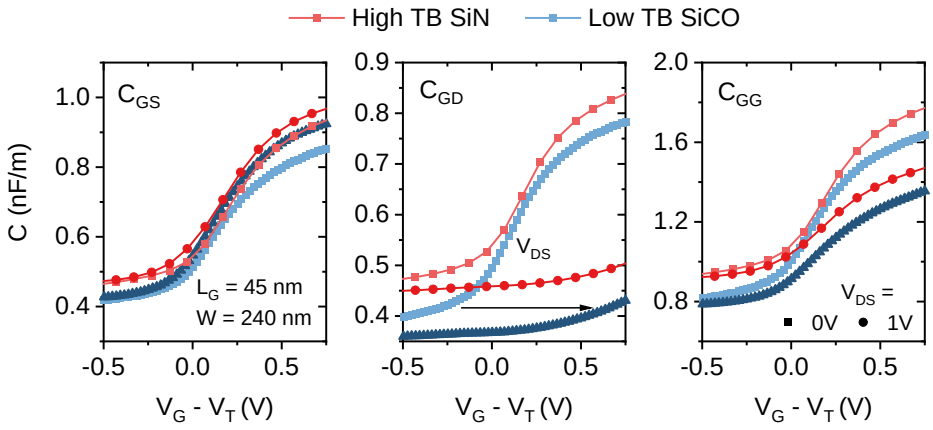


Figure 4-21 Device capacitive components versus V_G and V_D bias for SiN- HTB and SiCO-LTB devices. (left) Gate to source capacitance (C_{GS}). (middle) Gate to drain capacitance (C_{GD}). (right) Total gate capacitance (C_{GG}).

The effects of the reduction of parasitic capacitance on the transistor FoM are presented in **Erreur ! Source du renvoi introuvable.**. The LTB transistor shows much better values of F_T. Obviously, as the transconductance measured on both type of devices is very similar (see **Erreur ! Source du renvoi introuvable.**-left), this improvement results from the reduction of the parasitic component of the gate capacitance. The influence of both parameters is clearer on the maximum of F_T, when G_M is plotted against C_{GG} in saturation (see in **Erreur ! Source du renvoi introuvable.**-right,). The benefit on F_T due to the replacement of SiN by SiCO is about 10 GHz for a 45nm gate length device and becomes even higher when the gate length is further scaled (the spacers play a bigger role on the total gate capacitance).

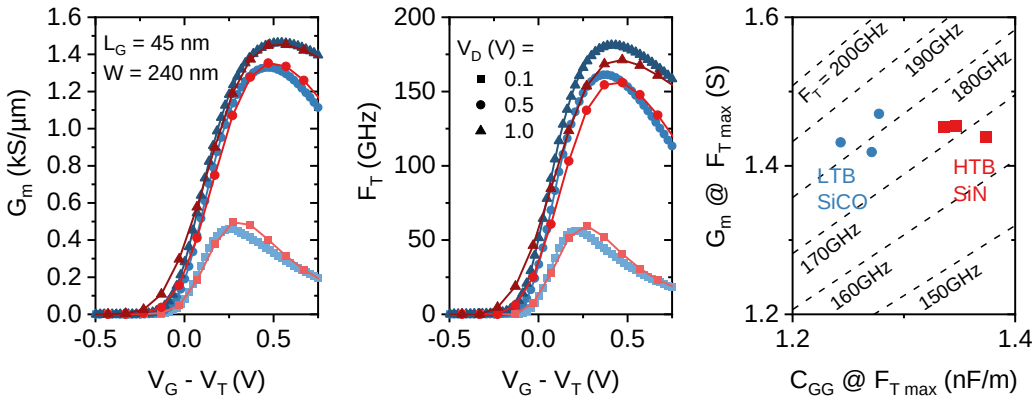


Figure 4-22(left) Transconductance and (middle) Cut-off Frequency of high and low thermal budget devices against gate and drain bias. (right) Transconductance vs Gate capacitance trade-off highlighting the impact of the spacer permittivity reduction on the cut-off frequency.

4.5 CAPACITANCE AND RESISTANCE TRADE-OFFS

4.5.1 Spacer thickness

From the previous analysis, it is clear that reducing the spacer permittivity allows boosting the device FoMs. However, considering the dielectric constant of the spacer is not sufficient for a complete optimization of the device RF FOM. The thickness of the material must also be taken into account. Basically the process flow of transistor fixes a given range of possible thickness for the spacer, for which the spacer actually plays its role of insulator between Gate and Source/Drain. Yet, changing the spacer thickness also modifies the access resistance of the device an in turn its performance. That is why in **Erreur ! Source du renvoi introuvable.** we propose to investigate the trade-off dielectric constant vs thickness on device RF FOMs This trade-off is established from TCAD simulations. For the study, the impact of the spacer permittivity (4.5 and 7) and thickness (5 nm to 40 mn) on RF FOMs of 30-nm NMOS FDSOI are assessed.

For a sake of simplicity, it is considered here that the source and drain are highly doped ($10^{20}\,cm^{-3}$), and that the junction is abrupt with overlap equal to zero. Considering a more complex junction profile or a different overlap would not significantly change the results of the study. The box and silicon film thickness are equal to 25 nm and 7 nm, respectively, which correspond to the expected dimensions from the 28-nm FD-SOI technology. The transport models used for this simulation take into account the degradation of mobility coming from doping, vertical electrical fields and saturation velocity. Furthermore, no strain is considered in these simulations.

G_M and C_{GG} characteristics for the low permittivity variants are reported in **Erreur ! Source du renvoi introuvable.**-left. As expected, by increasing the spacer thickness, the source and drain epitaxies and contacts are pushed away from the gate, resulting in a substantial reduction in total gate capacitance. However, increasing the spacer thickness also increases the length of the source and drain conductive path, and degrades the access resistance R_a and transconductance gain accordingly.

The result of this $R_a - C_{PAR}$ trade-off is highlighted in **Erreur ! Source du renvoi introuvable.**-right. The values G_M and C_{GG} are extracted at the maximum of F_T and plotted for each variant. The F_T contour lines are calculated using eq. 4-1. For

this particular junction profile and for both values of spacer permittivity, the reduction of parasitic capacitance clearly outweighs the degradation of the access resistance when the spacer thickness is increased from 5 nm to 30 nm. The simulation results show that the value of F_T improves by more than 60 GHz with thicker spacers, and can be even be further improved by 50 GHz by switching the spacer material from SiN to SiCO.

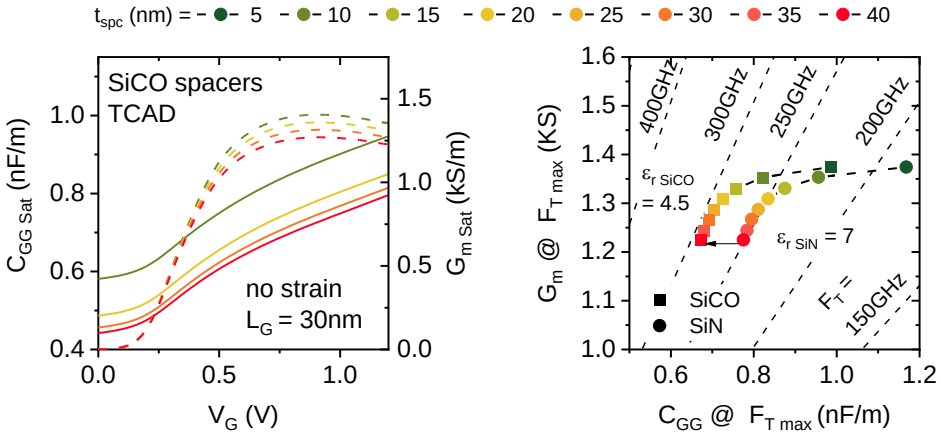


Figure 4-23 Comparative study of the thickness and permittivity of spacers using TCAD simulations. The spacer thickness varies between 5nm and 40nm while the doping density under the spacer is kept constant and equal to $10^{20}\ cm^{-3}$. (left) C_{GG} and G_M over V_G obtained from the simulation with SiCO spacers. (right) C_{GG} vs G_M at the maximum of F_T for each spacer thickness and two different materials with permittivity values of $SICO\varepsilon r = 4.5$ and $SIN\varepsilon r = 7$.

The previous results show that increasing the spacer thickness is suitable for high frequency application. However, the simulations are performed on transistors with highly doped source and drain. These doping features for the access are not practically achievable with thick spacers without an extension-first integration scheme.

Considering this issue, new TCAD simulations were performed with a lower density of dopants in the access ($N_D = 10^{19}\ cm^{-3}$), and a spacer permittivity fixed at 4.5. The results are compared to the previous simulations in **Erreur ! Source du renvoi introuvable.**. **Erreur ! Source du renvoi introuvable.**-left firstly reports the G_M vs. C_{GG} trade-offs. Reducing the dopant density in the access region increases the access resistance R_a and, degrades G_M accordingly. However, the drop on G_M also comes up with a reduction of C_{GG}. Basically, reducing the dopants density in the access region also tends to increase the space charge region of the drain junction in saturation regime compared to a highly doped drain. This significantly reduces the inner fringe gate to drain capacitance, and, in turn, the total gate capacitance C_{GG}. Nevertheless, the transconductance drop is generally stronger than the reduction of the gate

capacitance, inducing an overall reduction of the maximum cut-off frequency for the real device with lower density of dopants in the access.

Furthermore, the unique relationship between G_M vs. C_{GG} trade-off and spacer thickness is almost not affected by the profile of the junction. The cut-off frequency continues to follow the variation of spacer thickness, despite the degradation of the access resistance. The only difference is that the optimum thickness of the spacer goes from 35 nm to 25 nm when the density of dopants in the access decreases from $10^{20}\ cm^{-3}$ to $10^{19}\ cm^{-3}$.

Interestingly, the impact of lower LDD doping on F_{MAX} is not the same than on F_T . As F_{MAX} has a stronger dependence on the gate to drain capacitance than F_T, lowering the doping density is actually much more beneficial for F_{MAX}. **Erreur ! Source du renvoi introuvable.**-right presents the values of C_{GD} and F_T extracted at the maximum of F_{MAX}. F_{MAX} contour lines are estimated using eq. 4-2 for which G_{DS} contribution is considered as negligible. The gate resistance is generically fixed at 1 kΩ ohm for each finger. The trade-off on F_{MAX} is clearly different. Despite the huge drop on F_T, the dependence on the inverse of $C_{GD}{}^{-1/2}$ is sufficient to increase the maximum value of F_{MAX} when the density of dopants from source and drain is reduced.

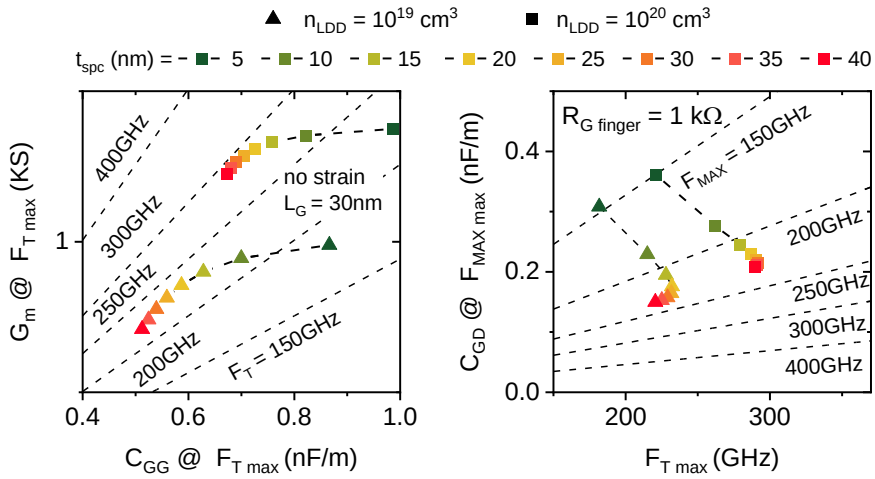


Figure 4-24 Effects of doping density in the access region and spacer thickness on RF FOMs. Two doping conditions are compared, $10^{20}\ cm^{-3}$ and $10^{19}\ cm^{-3}$. (left) C_{GG} vs G_M at the maximum of F_T for each variant. (right) F_T vs C_{GD} at the maximum of F_{MAX} for each variant. F_{MAX} contour are obtained from eq. 4-2 where the G_{DS} component is considered negligible.

The increase in F_{MAX} for low-doped access is highly beneficial for power amplifier applications because of the high V_{DD} sweep seen by the transistor under large RF signal operation. Basically highly doped and sharp junctions generally leads to

lower values of drain-source breakdown voltage as well as poorer hot-carrier reliability. By reducing the LDD doping density, the device has an improved reliability, a bigger value of F_{MAX} and a reduced power consumption. Moreover, for power amplifier circuits, F_{MAX} FOMs prevails over F_T. This is due to the fact that the amplification function is made in power rather than in current. Therefore, there is an obvious benefit on the junction engineering for further improving F_{MAX}. Because the benefit on F_{MAX} mainly arises from the gate to drain depletion capacitance, it is possible to optimize the transistor with a double implant under the spacer. Near the junction a low-doped LDD maximizes the depletion thickness at the desired drain bias, while a high-doped HDD minimizes the access resistance at the non-depleted region.

4.5.2 Non-symmetrical source and drain junction profiles

From the previous results, the ideal scenario would be to reduce source to drain capacitance, while keeping the same access resistance. Considering that the access resistance from the source has the strongest impact on the degradation of the transconductance [thor], this could be partially achieved by fabricating MOSFET with asymmetrical source and drain junction profiles. By keeping a high-doped source and low-doped drain, the source access resistance remains unaltered, while the depletion on the drain could reduce the C_{GD} capacitance in saturation regime.

These effects are visible in **Erreur ! Source du renvoi introuvable.** that shows the results of TCAD simulations performed with asymmetrical devices. The source doping density is kept constant at $10^{20}\ cm^{-3}$, while the dopant density in the region beneath the spacer of the drain varies between $10^{18}\ cm^{-3}$ and $10^{20}\ cm^{-3}$. The thickness of the spacer is near its optimal value of 30 nm and the permittivity of the spacer is 4.5. By reducing the doping density nearby the drain alone, the degradation of G_M is almost suppressed. The G_M of all the asymmetrical devices is greater than the one of the symmetrical device at any value of V_G, whereas the C_{GD} hardly changes . On the other hand, the C_{GS} increases when the drain LDD density decreases. This is explained by a shift of the channel pinch-off position towards the drain, that is translated into an increase of the effective channel length on saturation. This shift increases the drain depletion and results in a bigger gate to channel capacitance component in C_{GS}. The increase of C_{GS} out-weights the reduction of C_{GD} and the total

capacitance C_{GG} increases in saturation regime near the maximum of transconductance.

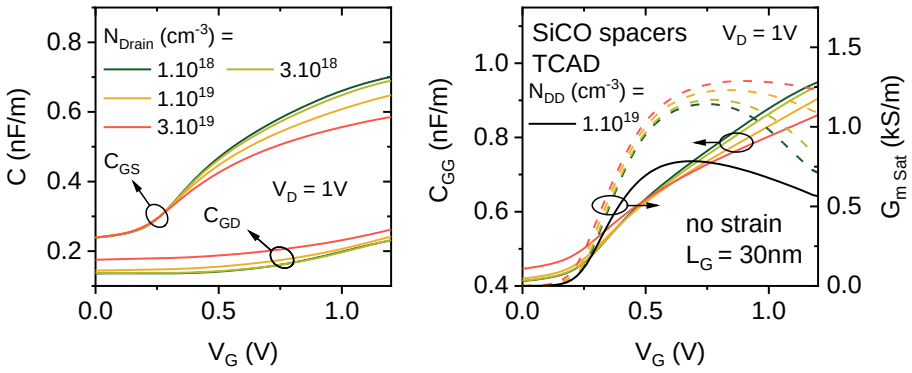


Figure 4-25 Effects of asymmetric source and drain doping on RF FoM. The density of dopants at the region underneath the spacer of the drain varies between $10^{18}\ cm^{-3}$ and $10^{20}\ cm^{-3}$ while the source is kept constant at $10^{20}\ cm^{-3}$. (left) C_{GD} and C_{GS} as function of V_G in saturation for each doping variant. (right) C_{GG} and G_M. Black curve corresponds to a symmetrical device with LDD density of $10^{19}\ cm^{-3}$.

When LDD doping is reduced in the asymmetrical devices, G_M decreases whereas C_{GG} increases, both contributing to lower F_T. In the worst-case, the maximum value of F_T drops of about 35 GHz. However, the effects on F_{MAX} still are different. The strong reduction of C_{GD} overcomes the F_T degradation seen formerly, leading to an increase of F_{MAX} of 25 GHz. Therefore, this optimization is actually highly recommended for power amplifier transistors for the same reason than previously i.e. F_{MAX} is more important than F_T for power transistor.

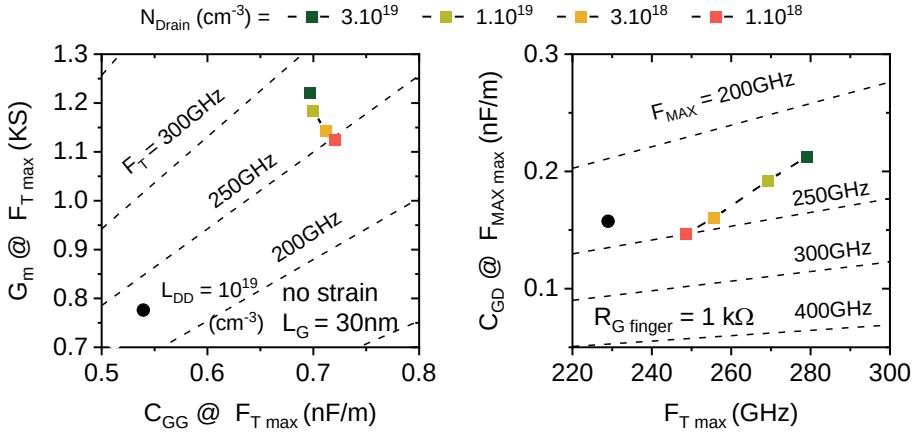


Figure 4-26 FoMs obtained from each drain LDD variant. Black-round point correspond to a symmetric device with LDD density of $10^{19}\ cm^{-3}$. (left) C_{GG} vs G_M at the maximum of F_T. (right) F_T vs C_{GD} at the maximum of F_{MAX}.

But more interestingly is the benchmark of these asymmetrical devices against the symmetrical transistor with LDD of 10^{19}cm^{-2}. As depicted in Figure 4-26, the F_{max}

optimization, in these devices, can be made without sacrifying F_T, unlike in symmetrical MOSFET. This is because the asymmetric device still has a low access resistance on the source terminal.

4.5.3 Overlap position

A different trade-off is related to the overlap position of source/drain junction. This effect can be qualitatively studied on the variants of S/D implantation presented in chapter 2. **Erreur ! Source du renvoi introuvable.**-left compares the curves of transconductance and capacitance of 28-nm S1 and S4 devices. S1 is the most underlapped transistor (ov = -5 nm) while S1 is the lesser one (ov = -2 nm). Like the drain current, the transconductance highly increases when underlap is reduced, because of the strong reduction of the access resistance. Regarding the total gate capacitance, in the off state it reduces when the underlap rises, mostly because the reduction of the inner field capacitance. In on state, the benefit yet tends to disappear.

The impact on F_T is presented in **Erreur ! Source du renvoi introuvable.**-right. Since no RF devices were available for characterization on the studied wafers, F_T is directly calculated from DC measurements using eq. 4-1. This model is enough reliable and accurate to assess F_T even if it does not account for self-heating that may induce loss in DC transconductance: former RF measurements show a very good agreement with modeling. Basically the more overlapped is the device, the higher is the F_T. This is due to the fact that the G_M increase with overlap is much more important than the increase in parasitic capacitances . The effect on F_T is so strong that the same trend should also be expected for F_{MAX}, even if the parasitic capacitances have more impact on this FoM.

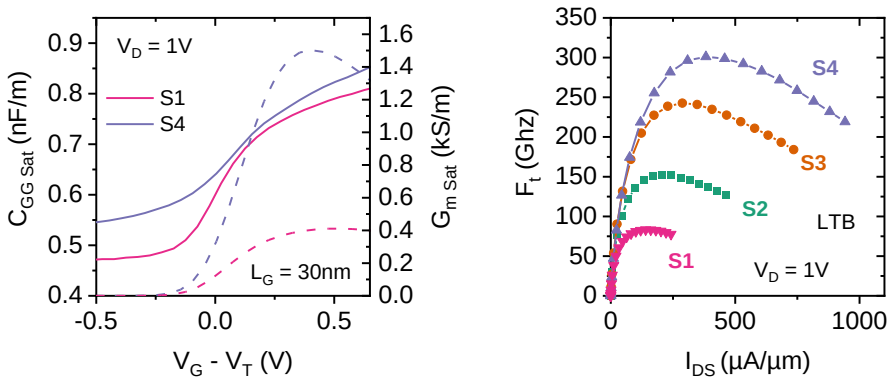


Figure 4-27(left) C_{GG} and G_M extracted from DC measurements of LTB S1 and S2 devices (right) F_T estimated from 28nm DC C_{GG} and G_M.

It should be noted that underlapping the junction results in a more significant degradation of transconductance than reducing the density of dopants while maintaining the overlap constant. This outcome is largely due to the exponential rise in resistivity in non-doped areas. Additionally, the un-doped region is more vulnerable to the electrostatic repulsion effect of traps on the spacers

Erreur ! Source du renvoi introuvable. compares the RF figures of merit of HTB devices with the ones of our best LTB devices. Both devices have the same gate length of 30 nm but different architecture. In particular, the spacer of the HTB transistor is a 25 nm SiN layer while the one of LTB device is a 7nm SiCO material. In Figure 4-28-left, we can see that both devices present similar values of F_T over overdrive bias, although their architecture is rather different. This effect is explained in **Erreur ! Source du renvoi introuvable.**-middle by looking at the values of G_M and C_{GG} extracted at the max of F_T. While C_{GG} is smaller for HTB devices due to the thick spacers, the LTB device compensates with a greater transconductance. Moreover, the fringe capacitance is reduced in LTB compared to LTB since the low permittivity of the spacer compensates its bigger thickness. Both effects compensate each other, giving similar values of F_T for both devices.

Another interesting effect that can be evidenced from this figure is that optimizing a transistor for DC performance does not necessary lead to the best optimization for RF applications. Indeed, while it would be compelling to maximize the saturated drain current on saturation for digital applications, the trade-off between capacitance and current imposes a different optimization at high frequencies. Indeed the optimum F_T and F_{MAX} are not necessarily obtained with the highest drain current density. Here increasing the spacer thickness to reduced fringe capacitance largely compensate the reduction of saturation current at fixed V_{DD} seen elsewhere.

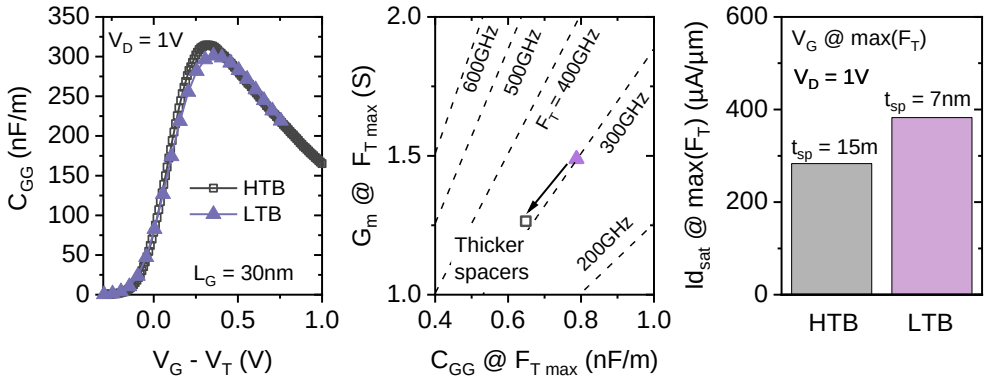


Figure 4-28(left) Comparison between estimated LTB F_T and measured HTB F_T as function of Vg. Both devices have 28nm gate length. (middle) C_{GG} vs G_M at the maximum of F_T for both devices. F_T contours are calculated using eq. 4-2. (right) Drain current in saturation regime at the max of F_T for both variants.

Finally, the previous results lead to guidelines for further improvement of the transistor at high frequency. By increasing the LTB devices spacer thickness while maintaining the overlap, it could be possible to further increase the value of F_T as presented in **Erreur ! Source du renvoi introuvable.**. In addition to the better high frequency FoMs, the device would also consume less power because of the lower drain current of this device at same bias. However, it would be impossible to keep the same overlap position, while increasing the spacer thickness on a low temperature extension-last integration scheme as explained in chapter 2.

From the conclusions drawn in the previous chapters, we propose to circumvent this issue by performing an extension-first LDD doping with an activation level near $10^{20}\ cm^{-3}$ using a non-amorphizing implantation before the integration of the spacers. In addition to the reduction of spacer traps with a hydrogen anneal, the high density of dopants would be enough to screen the electrostatic repulsive effects from the SiCO traps. From a successful extension first integration with a 30-nm SiCO spacer and similar junction position, it could be expected an improvement of near 50 GHz on F_T for a total of 350 GHz. This fabrication scheme could be conceivable for a high frequency dedicated top-layer for RF transceiver SoCs designed for mm-wave applications.

4.6 NANO-SECOND LASER ANNEAL ON GATE RESISTANCE

For the HTB devices, the gate polysilicon is classically doped and activated by the same annealing used for the activation of source and drain. However, this high temperature process cannot be performed on an LTB fabrication flow. Moreover, the gate polysilicon does not contain a crystalline seed. Therefore, the activation by Solid phase epitaxy recrystallization (SPER) recrystallization followed by the thermal activation with a 500°C anneal is not possible either for the gate case. To solve this issue, we used an approach based on the crystallization of the polysilicon gate with UV nanoseconds laser anneal (UV-NLA) [13]. The laser anneal is performed after the deposition of a doped in-situ polysilicon, and before etching the gate stack. This allows a homogenous heat absorption from the polysilicon layer and reduce variability between devices with different geometries. The very short periodic pulses used by UV-NLA allows the transfer of heat to the top tier while keeping the bottom layers below 500°C. The laser anneal is followed by a CMP step to polish the polysilicon surface.

Gate resistance depends on several gate stack components [11] as presented in **Erreur ! Source du renvoi introuvable.**-left. Each component contributes to the horizontal, vertical or contact resistance from eq. 4-6. R_{ct} refers to the BEOL metallic contacts and contact resistance between tungsten and silicide. R_H is the silicide resistance that solely contributes to the horizontal resistance. Finally, the vertical resistance R_V is the sum of polysilicon, TiN bulk resistances, and of the metal-semiconductor interface resistance associated to a Schottky barrier obtained from the between silicide and polysilicon.

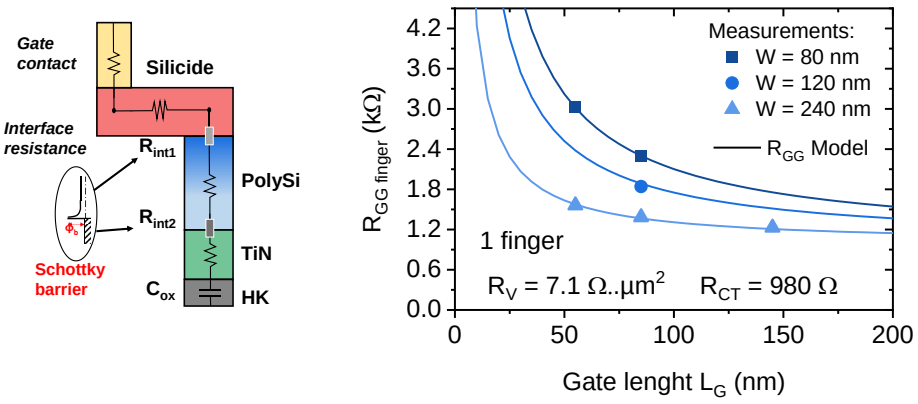


Figure 4-29(left) sketch illustrating the different components of the gate resistance (right) Measurement of gate resistance for different geometries and extraction of vertical and contact component with the fit of the model of eq. 4-6.

Erreur ! Source du renvoi introuvable.-right presents the measured gate resistance for a NMOS LTB device with different values of gate length and width. The model of eq. 4-6 is used to fit the data and to extract the vertical and contact resistance. The horizontal resistance itself cannot directly be extracted from this set of measurements. To assess it, larger devices are required for which this component is not negligible compared to the other resistances. Nevertheless, the device presents a vertical resistance of $7\,\Omega.\mu m^2$ and a contact resistance of 980 Ω. The vertical component is remarkably in great agreement with what is expected from the literature [ref]. This result indicates that the activation of dopants at the gate with laser anneal was successful and demonstrate that UV-NLA is a reliable solution for the polysilicon dopants activation on LTB devices.

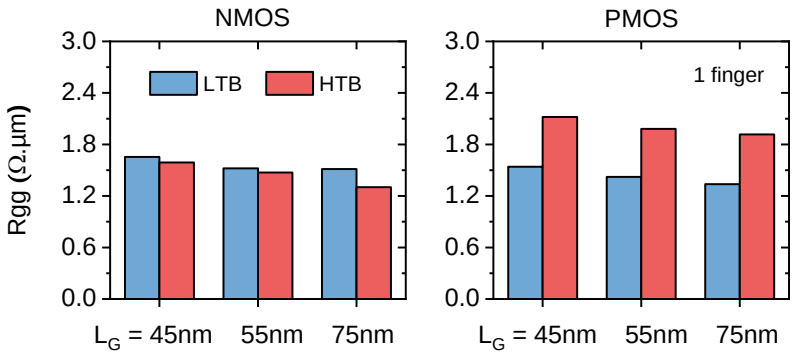


Figure 4-30 Comparison of gate resistance obtained from LTB and HTB devices. The resistance is extracted for three gate lengths on NMOS and pMOS devices.

The gate resistance from CMOS devices fabricated with an LTB process is compared to HTB with same dimensions in **Erreur ! Source du renvoi introuvable.**. The gate resistance from NMOS transistors is similar to the high temperature process for every gate length. But the comparison is not the same with PMOS. Because the gate is doped in-situ, it can only contain one type of impurities. Therefore, the polysilicon part of both gates are doped with phosphorous and are n-type. On the other hand, the gate used in the high temperature PMOS is doped with p-type impurities that are commonly known to have greater resistivity [Hideo Muro]. It also explains why both n and p-type devices have similar values of gate resistance with the LTB process.

4.7 STATE OF THE ART RF FOMS OF LTB TRANSISTORS

In this part, we report the state-of-the-art RF performance of the FD-SOI transistor fabricated in LETI within a maximum thermal budget of 500°C. Following the guidelines proposed all along this manuscript, we carefully optimize each key process step that affects the high frequency behavior of our device i.e. the channel mobility, gate resistance, parasitic capacitance and access resistance. The proposed solutions are compliant with the Low Thermal Budget required for the integration of these transistors in a 3DSI technology. Results are shown in **Erreur ! Source du renvoi introuvable.** & Figure 4-32 for LTB NMOS and PMOS 45-nm FD-SOI devices.

Firstly, NMOS FDSOI devices are considered. They feature F_T of 180 GHz and F_{MAX} of 240 GHz at $V_{DD} = 1V$. These values are the best ever reported for Silicon-based transistors fabricated within a low thermal budget <500°C. On the other hand PMOS devices, feature F_T = 105 GHz and F_{MAX} = 175 GHz at $V_{DD} = -1V$. Again, these values are at the state-of-the-art for a Coolcube technology.

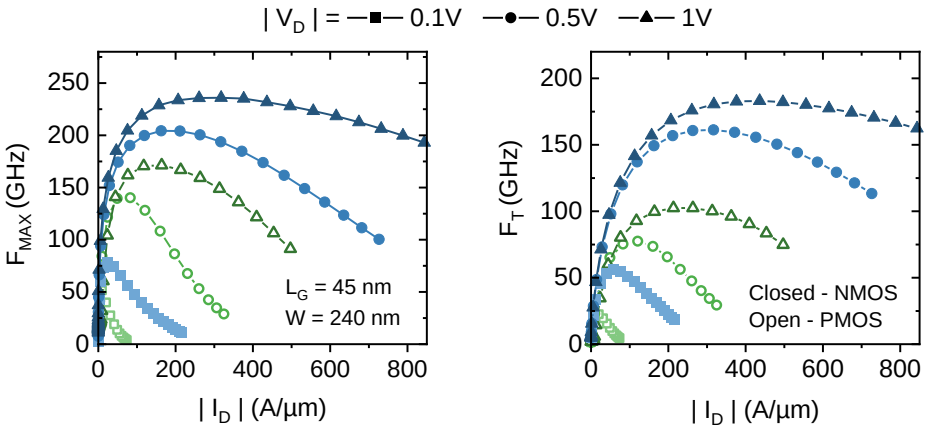


Figure 4-31 LTB FoM of NMOS and PMOS devices with $L_G = 45\text{nm}$ and $W = 240\text{nm}$ as the function of drain current at different drain bias. (left) F_{MAX}. (right) F_T.

The RF performance of these LTB silicon-based NMOS and PMOS devices are then benchmarked to their High Thermal Budget counterparts in **Erreur ! Source du renvoi introuvable.**. For a fair comparison, data are plotted this time against the reverse of L_G. As a reminder, F_T is expected to vary linearly with L_G^{-1} while F_{max} is roughly proportional to $L_G^{-1/2}$. The key result is that the record RF performance for our LTB n and p-type devices is actually directly comparable to the best values found in the literature for HTB Si transistors. Furthermore, both type of devices exhibit the

similar dependence with gate length than the one expected by simple modelling (solid lines). This demonstrates that the underlying physics explaining the RF behaviour of the device in unchanged in our LTB transistors.

Basically the great results on F_T roughly come from the combination of low permittivity SiCO spacers, low access resistance and tensile strain from the CESL. It proves that high performance devices could be fabricated on dedicated high frequency top-layers. In addition, with further optimization of the spacer thickness, our LTB 28-nm FDSOI devices could present FoMs close to the record ones for silicon CMOS technology. Furthermore, our LTB devices shows equivalent F_{MAX} values for NMOS, and even superior for PMOS at same gate length. These good results in our LTB devices, compared to common HTB transistor, can be explained by

1. the use of the low-k spacer, which reduces C_{GD} for both NMOS and PMOS
2. the lower gate resistance in the specific case of the PMOS that is explained by the replacement, in our technology, of a standard p-implanted polysilicon gate by an in-situ n-type polysilicon gate.

Note that, although our LTB device already shows great performances, we can undoubtedly further improve them by optimizing the design of transistor. Indeed, the LTB devices tested here are designed with only one gate contact by finger. With a two-contact gate design [ref], we estimate that the gate resistance can be reduced by a factor 2, improving the values of F_{MAX} by nearly 40%. As a conclusion, this LTB technology has a great potential for RF application, all the more that the low temperature process authorizes the integration of these transistor as a top tier of a 3DSI technology.

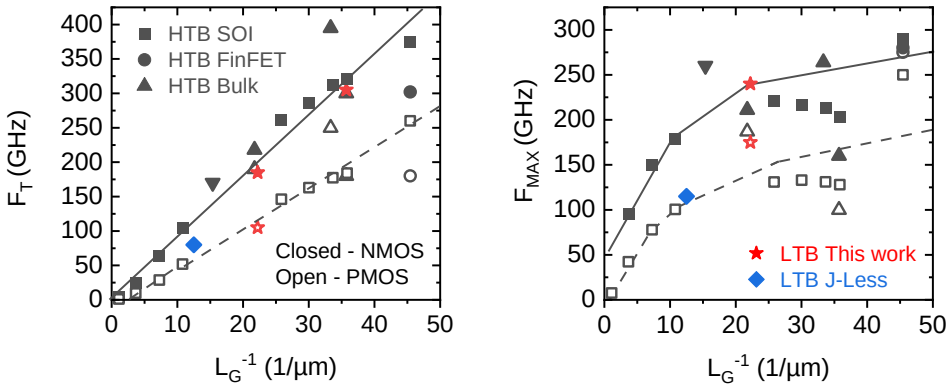


Figure 4-32 Comparative RF performance of LTB devices fabricated at LETI against the state-of-the-art silicon-based NMOS & PMOS devices. F_T and F_{MAX} of LETI devices are comparable to HTB devices and present a similar behavior over $L_G{}^{-1}$(left) max of F_{MAX}. (right) max of F_T. Note that the F_T

of the 30nm LTB device is not directly measured in that case, but it is estimated from low frequency capacitance and DC current measurements.

4.8 CONCLUSION

In this chapter, we presented a revisited method for the extraction of the channel effective mobility using two devices with different gate lengths. In addition, a method to extract the effective mobility accurately using S-parameters measurements was proposed and validated against measurements made on traditional high temperature MOSFET transistors. The new method is able to extract the mobility of nanometer devices over the full range of inversion charge. From the extracted mobility, it was possible to study in details the variation of mobility along the device channel. Using this revisited method, we also analyse the benefits brought of a high quality oxide o on carriers scattering and effective mobility, as well as the boost effect of tensile CSEL strain on mobility and reliability of short channel LTB devices.

Next, we studied how some key steps of the low temperature process can affect the RF performance of our device. We showed that the fringe parasitic capacitance, and thereby cut-off frequency, is improved in our technology because of the use of low-k SiCO spacers. In addition, using TCAD simulations, we explored the trade-offs originated from the optimization the spacer thickness, source and drain doping density and junction position. It was concluded that a 25 to 35nm thick spacer maximizes the F_T and F_{MAX} figures of merit because of the trade-off between the fringe parasitic capacitances and the access resistance. Finaly, we explored the effects of an asymmetric low-doped drain transistor for improved maximum oscillation frequency and longer reliability for Power amplifiers. We also presented a solution to activate dopants of the polysilicon gate in a LTB process, with a UV nanoseconds laser anneal. It was shown that the laser anneal allows an high level of activation so that the gate resistance of our LTB devices is comparable to the one measured on high thermal budget references.

Finally, we benchmark F_T and F_{MAX} of the LTB devices fabricated at CEA with the state of the art HTB silicon MOSFETs. It is highlighted that low temperature process flow can be used to fabricate high performance RF transistors with similar FoMs to high temperature silicon MOSFET counterparts.

Most of the results from this chapter were presented at the VLSI Symposium of 2022 and in the issue of July 2021 of the Transactions on Electron Devices (TED) journal.

4.9 REFERENCES

[20] D. Dutoit et al., "How 3D integration technologies enable advanced compute node for Exascale-level High Performance Computing" 2020 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2020, pp. 311 - 314.

[21] A. O. Watanabe, M. Ali, S. Y. B. Sayeed, R. R. Tummala and P. M. Raj, "A Review of 5G Front-End Systems Package Integration," in IEEE Transactions on Components, Packaging and Manufacturing Technology, doi: 10.1109/TCPMT.2020.3041412.

[22] L. Hutin et al., "Gate reflectometry for probing charge and spin states in linear Si MOS split-gate arrays," 2019 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2019, pp. 37.7.1-37.7.4 , doi: 10.1109/IEDM19573.2019.8993580.

[23] C. Thomas et al., "Die-to-Wafer 3D Interconnections Operating at Sub-Kelvin Temperatures for Quantum Computation," 2020 IEEE 8th Electronics System-Integration Technology Conference (ESTC), Tønsberg, Vestfold, Norway, 2020, pp. 1-7, doi: 10.1109/ESTC48849.2020.9229657

[24] C. Fenouillet-Beranger et al., "A review of the full 500°C low temperature technological modules development for high performance and reliable 3D Sequential Integration," 2019 Electron Devices Technology and Manufacturing Conference (EDTM), doi: 10.1109/EDTM.2019.8731192.

[25] C. M. V. Lu et al., "Key process steps for high performance and reliable 3D Sequential Integration," *2017 Symposium on VLSI Technology*, Kyoto, 2017, pp. T226-T227, doi: 10.23919/VLSIT.2017.7998181.

[26] J. Micout et al., "Towards 500°C SPER activated devices for 3D sequential integration," 2017 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S), Burlingame, CA, 2017, pp. 1-2, doi: 10.1109/S3S.2017.8309220.

[27] A. Tsiara et al., "Performance and Reliability of a Fully Integrated 3D Sequential Technology," 2018 IEEE Symposium on VLSI Technology, Honolulu, HI, 2018, pp. 75-76 , doi: 10.1109/VLSIT.2018.8510625.

[28] C. Cavalcante et al., "28nm FDSOI CMOS technology (FEOL and BEOL) thermal stability for 3D Sequential Integration: yield and reliability analysis," 2020 IEEE Symposium on VLSI Technology, Honolulu, HI, USA, 2020, pp. 1-2, doi: 10.1109/VLSITechnology18217.2020.9265075.

[29] C. Fenouillet-Beranger et al., "First demonstration of low temperature (≤ 500°C) CMOS devices featuring functional RO and SRAM bitcells toward 3D VLSI integration," 2020 IEEE Symposium on VLSI Technology, Honolulu, HI, USA, 2020, pp. 1-2, doi: 10.1109/VLSITechnology18217.2020.9265092.

[30] X. Garros et al., "RF Performance of a Fully Integrated 3D Sequential Technology," 2019 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2019, pp. 25.1.1-25.1.4, doi: 10.1109/IEDM19573.2019.8993512.

[31] D. Benoit et al., "Interest of SiCO low k=4.5 spacer deposited at low temperature (400°C) in the perspective of 3D VLSI integration," 2015 IEEE International Electron Devices Meeting (IEDM), Washington, DC, 2015, pp. 8.6.1-8.6.4, doi: 10.1109/IEDM.2015.7409656.

[32] L. Brunet et al., "Breakthroughs in 3D Sequential technology," 2018 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2018, pp. 7.2.1-7.2.4 , doi: 10.1109/IEDM.2018.8614653.

[33] Tsividis, Y. (1999), “Operation and Modeling of the MOS Transistor”. 2nd Edition, McGraw-Hill, New York, pp 534 – 547.

[34] O. Weber et al., "14nm FDSOI upgraded device performance for ultra-low voltage operation," 2015 Symposium on VLSI Technology (VLSI Technology), Kyoto, 2015, pp. T168-T169, doi: 10.1109/VLSIT.2015.7223664.

[35] L. Pasini et al., "High performance CMOS FDSOI devices activated at low temperature," 2016 IEEE Symposium on VLSI Technology, Honolulu, HI, 2016, pp. 1-2 , doi: 10.1109/VLSIT.2016.7573407.

[36] T. M. Frutuoso et al., "Impact of spacer interface charges on performance and reliability of low temperature transistors for 3D sequential integration," 2021 IEEE International Reliability Physics Symposium (IRPS), 2021, pp. 1-5, doi: 10.1109/IRPS46558.2021.9405107.

[37] A. Tsiara et al., "Performance and Reliability of a Fully Integrated 3D Sequential Technology," 2018 IEEE Symposium on VLSI Technology, Honolulu, HI, 2018, pp. 75-76 , doi: 10.1109/VLSIT.2018.8510625.

[38] J. Franco et al., "BTI Reliability Improvement Strategies in Low Thermal Budget Gate Stacks for 3D Sequential Integration," 2018 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2018, pp. 34.2.1-34.2.4 , doi: 10.1109/IEDM.2018.8614559

[39] S. -H. Kim et al., "Heterogeneous Integration Toward a Monolithic 3-D Chip Enabled by III–V and Ge Materials," in IEEE Journal of the Electron Devices Society, vol. 6, pp. 579-587, 2018, doi: 10.1109/JEDS.2018.2802840.

[40] W. Rachmady et al., "300mm Heterogeneous 3D Integration of Record Performance Layer Transfer Germanium PMOS with Silicon NMOS for Low Power High Performance Logic Applications," 2019 IEEE International Electron Devices Meeting (IEDM), 2019, pp. 29.7.1-29.7.4, doi: 10.1109/IEDM19573.2019.8993626.

[41] H. W. Then et al., "3D heterogeneous integration of high performance high-K metal gate GaN NMOS and Si PMOS transistors on 300mm high-resistivity Si substrate for energy-efficient and compact power delivery, RF (5G and beyond) and SoC applications," 2019 IEEE International Electron Devices Meeting (IEDM), 2019, pp. 17.3.1-17.3.4, doi: 10.1109/IEDM19573.2019.8993583.

[42] J. Raskin, "SOI technology pushes the limits of CMOS for RF applications," 2016 IEEE 16th Topical Meeting on Silicon Monolithic Integrated Circuits in RF Systems (SiRF), Austin, TX, 2016, pp. 17-20 , doi: 10.1109/SIRF.2016.7445456.

[43] P. Sideris et al., "Inter-tier Dynamic Coupling and RF Crosstalk in 3D Sequential Integration," 2019 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2019, pp. 3.4.1-3.4.4 , doi: 10.1109/IEDM19573.2019.8993493.

[44] P. Sideris, L. Brunet, G. Sicard, P. Batude and C. Theodorou, "Impact of Inter-Tier Coupling on Static and Noise Performance in 3D Sequential Integration Technology," Solid-State Electronics, vol 168, 2020, doi: 10.1016/j.sse.2019.107715.

[45] A. Vandooren et al., "Buried metal line compatible with 3D sequential integration for top tier planar devices dynamic Vth tuning and RF shielding applications," 2019 Symposium on VLSI Technology, Kyoto, Japan, 2019, pp. T56-T57, doi: 10.23919/VLSIT.2019.8776490.

[46] P. Batude, IEDM'17
[47] L. Brunet, IEDM'18
[48] A. Kumar, VLSI'17
[49] A. Vandooren, TED'18, pp. 5165
[50] D. Choudhury, IMS'10
[51] V. Deshpande, SSE, 2017, pp. 87
[52] V. Lu, VLSI'17,[8] C. Fenouillet-Beranger, VLSI'19,
[53] X. Garros IEDM'19,
[54] T. Mota Frutuoso IRPS'21,
[55] D. Fleury ICMTS'08, [
[56] A. Tsiara, [13] Z. Zhao ESSDERC'19
[57] C.-H. Jan IEDM'10,
[58] Samsung 28RF,
[59] P. VanDerVoorn VLSI'10
[60] R. Carte IEDM'16,
[61] H.-J. Lee IEDM'18

Chapter 5: Conclusion

The integration of 3D sequential technology has shown potential to overcome the scaling limitations of 2D architectures. However, in order to implement this technology, MOSFET integrated circuits need to undergo fabrication steps that often involve high temperatures of over 1000°C. The main challenge with 3D sequential integration is to develop a low temperature process where all steps can be performed within a maximum thermal budget of 500°C. The aim of this Ph.D. project was to investigate the effect of the low temperature steps developed at LETI and compare the performance of low thermal budget transistors with those processed using high thermal budget in the industry. Additionally, a study was conducted to evaluate the performance of the devices at high frequencies, which is necessary for assessing their suitability for use in radio frequency and millimeter wave applications. Nonetheless, to obtain a deeper understanding of the electrical effects of each process, new and more advanced characterization methods were developed as a result of this study. Those new methods can also be useful on the development of more advanced FDSOI MOSFET nodes processed with a High thermal budget.

Chapter one provides an introduction to the fundamental principles of 3D sequential integration, including a detailed overview of the differences between low-temperature steps and high-temperature steps. The chapter also outlines the specific fabrication steps utilized during the production of the 28 FDSOI node at low temperatures.

Chapter two explores the effects of the low temperature annealing on the activation of dopants under the spacers and the formation of the source and drain junction. To perform this analysis, a non-destructive characterization method called CV-AJP was proposed, which allows the extraction of the junction profile from gate capacitance measurements with high accuracy. This method was used to study the impact of drain and source implantation and activation on NMOS and PMOS transistor performance, leading to the identification of optimal implantation conditions for overlapped devices and a conclusion that a strong germanium implantation can degrade device performance. Additionally, the thermal solubility of impurities was found to be greater under the spacers due to the proximity to the oxides interface of

the spacers and the BOX. This, collusion, can enable the development of a more reliable extension first integration scheme device without the need for SPER recrystallization near the Junction. Finally, the effects of junction overlap on Hot Carriers degradation were studied, showing that the HC degradation at high stress is due to both electron trapping in the oxide spacers and channel/gate oxide degradation. The first contribution is reduced with overlapped devices, while the second is mainly controlled by the level of the drain current and increases with overlap.

Chapter three study the trapping mechanism on the SiCO material used for spacers of LETI. For this, two new fast capacitance measurement techniques where developed, namely CV_{pulse} and CV_{ramp}, which can be used to characterize traps and ageing of MOS capacitors. Two types of defects were identified: fast Si interface traps, which depend on the quality of the native oxide, and slow deep defects distributed in the bulk of the SiCO oxide. The study was furthered to understand the effects of forming gas and spike annealing on SiCO trap passivation and de-passivation. The results showed that a forming gas anneal was effective in passivating a large number of oxide and interface traps, while high-temperature annealing de-passivated the Si interface, generating a large density of interface traps while reducing the deep oxide trap density. However, the increase in interface trap density due to the thermal anneal could be partially recovered with a subsequent forming gas anneal. In addition, this chapter discusses the effects of interface traps in the spacer region on the performance and reliability of 30nm FDSOI low-temperature transistors through CV measurements and TCAD simulations. The presence of spacer charges induces the formation of a depleted zone below the spacers, which is responsible for increasing access resistance and abnormal V_T variation with device scaling. However, the presence of dopants near the defects attenuates the negative effects of the interface charges, and thus, this issue is detrimental only for underlapped channels.

Chapter four evaluate the performance of the devices at high frequencies. The effects of low-temperature process steps on RF Figures of merit are explored and compared to high temperature FDOSI devices. The low-k SiCO spacers improves fringe parasitic capacitance with has a beneficial effect on the cut-off frequency. In addition, the activation of gate dopants using UV nanoseconds laser anneal is show to produce values of gate resistance similar to the high temperature anneals. Trade-offs are examined for optimizing spacer thickness, source and drain doping density, and

junction position. An asymmetric low-doped drain transistor is proposed for improved maximum oscillation frequency and longer reliability for power amplifiers Finally, the F_T and F_{MAX} of the LTB devices are benchmarked against the state of the art HTB silicon MOSFETs, showing that low-temperature processes can produce high-performance RF transistors with similar FoMs to HTB counterparts.

List of publications

Fenouillet-Beranger , L. Brunet , P. Batude , L. Brevard , X. Garros , **T. M. Frutuoso** et al., "First Demonstration of Low Temperature (≤500°C) CMOS Devices Featuring Functional RO and SRAM Bitcells toward 3D VLSI Integration," VLSI 2020

T. M. Frutuoso et al., "Impact of spacer interface charges on performance and reliability of low temperature transistors for 3D sequential integration" IRPS 2021

T. M. Frutuoso et al., "RF Performance of Devices Processed in Low-Temperature Sequential Integration," TED 2021

E. Catapano, G.Ghibaudo, M. Cassé, **T. M. Frutuoso** et al., "Statistical and Electrical Modeling of FDSOI Four-Gate Qubit MOS Devices at Room Temperature," in JEDS 2021

R. Kom Kammeugne, C. Leroux**, T. M. Frutuoso**, et al., "Parasitic Capacitance Analysis in Short Channel MIS-HEMTs GaN", ESSDERC 2021

T. M. Frutuoso et al., "Record RF Performance (ft=180GHz and fmax=240GHz) of a FDSOI NMOS processed within a Low Thermal Budget for 3D Sequential Integration" VLSI 2021

L. Brunet, S. Reboh, T. Januel, X. Garros, **T. M. Frutuoso** et al., "Record Performance of 500°C Low-Temperature nMOSFETs for 3D Sequential Integration using a Smart CutTM Layer Transfer Module" VLSI 2021

P. Batude, O. Billoint, S. Thuries, P. Malinge, C. Fenouillet, A. Peizerat, G. Sicard, P. Vivet, S. Reboh, C. Cavalcante, L. Brunet, M. Ribotta, L. Brevard, X. Garros, T**. M. Frutuoso** et al, "3D sequential integration: applications and associated Key Enabling Modules (design & technology) ", IEDM 2022

T. M. Frutuoso et al., "Ultra-fast CV methods (< 10µs) for interface trap spectroscopy and BTI reliability characterization using MOS capacitors", IRPS 2022

T. M. Frutuoso et al., "Methodology for Active Junction Profile Extraction in thin film FD-SOI Enabling performance driver identification in 500°C devices for 3D sequential integration", VLSI 2022