Ultra-thin plasma nitrided oxide gate dielectrics for advanced MOS transistors

Florence Cubaynes

Composition of the Graduation Committee:

Chairman:

Secretary:

Promoters:
Prof. dr. J. Schmitz
Prof. dr. ir. K. de Meyer

Assistant Promotor:
Dr. ir. C. Salm

Internal Members:
Prof. dr. F. Kuper
Prof. dr. T. Mouthaan

External Member:
Prof. dr. U. Schwalke

Referee:
Dr. P. Woerlee

Title: Ultra-thin plasma nitrided oxide gate dielectrics for advanced MOS transistors

Author: Florence Cubaynes

Keywords:

ISBN ….
Printed in …. Country ….

Ultra-thin plasma nitrided oxide gate dielectrics for advanced MOS transistors

Dissertation

to obtain
the doctor's degree at the Universiteit of Twente,
on the authority f the rector magnificus
prof. dr. F.A. van Vught,
on account of the decision of the graduation committee,
to be publicly defended
on Thursday 24th of June at 16.45

by

FLORENCE CUBAYNES
born on August 17th, 1976
in Toulouse, France

This dissertation is approved by promoters:
Prof. dr. J. Schmitz
Prof. dr. ir. K. de Meyer

and assistant promoter
Dr. ir. C. Salm

"J'ai appris qu'une vie ne vaut rien, mais que rien ne vaut une vie"
A. Malraux

In honor of my uncle Charles

To Paul

Table of Contents

Chapter 1
Introduction

1.1 Scaling the MOSFET

1.1.1 Motivation

The semiconductor industry strongly relies on its ability to continuously reduce the vertical and lateral dimensions of the Silicon (Si) based metal-oxide-semiconductor field effect transistor (MOSFET). This scaling of MOSFETs and other devices is following the well-known Moore's law ([1], [2]). The motivation for this scaling is to increase the packing density resulting in greater integrated circuit functionality, performance, and reduction of overall cost.
The performance of a complementary MOS (CMOS) circuit can be characterized through the dynamic response of the transistors that constitute the circuit, called the switching time. When considering an ideal CMOS inverter circuit, the switching time (τ) is inversely proportional to the drive current of the N- and PMOS devices, as shown in Equation 1.1. Therefore, an increase of the drive current of the MOSFET results in an improvement of the circuit performance.

$$\tau = f(V_{DD}) \times (\tau_n + \tau_p) \tag{1.1}$$

with

$$\tau_n = \left(\frac{C_{ox} \times V_{DD}}{I_{ON}}\right)_n \text{ and } \tau_p = \left(\frac{C_{ox} \times V_{DD}}{I_{ON}}\right)_p$$

where
τ_n and τ_p are the switching time of the NMOS and PMOS transistors, respectively. V_{DD}, C_{ox} and I_{ON} are the supply voltage, the oxide capacitance and the drive current, respectively. f(V_{DD}) is a function that depends linearly on V_{DD}.

The reduction of the channel length and the gate dielectric thickness (inversely proportional to C_{ox}) of the MOSFET associated to the scaling, leads to an increase of the drive current (Equations 1.2 and 1.3). A long and short channel MOSFET in the saturation regime was assumed in Equations 1.2 and 1.3, respectively.

$$I_{Dsat} = \frac{W}{2 \times L_G} \times \mu_{eff} \times C_{ox} \times (V_G - V_T)^2 \quad (1.2)$$

$$I_{Dsat} = W \times C_{ox} \times \upsilon_{sat} \times (V_G - V_T) \quad (1.3)$$

where, I_{DS}, μ_{eff} and C_{ox} are respectively the drain current, the effective channel mobility and the oxide capacitance (inversely proportional to the gate dielectric thickness). υ_{sat} is the velocity saturation. W is the channel width and L_G the channel length. V_{GS} and V_T are respectively the gate to substrate voltage drop and the threshold voltage.

Various scaling methodologies have been used through the past decades such as constant electric field or constant voltage scaling. Constant electric field scaling method reduces all dimensions by a common factor, K. However, while the drive current increases by K, the supply voltage is reduced which results to an incompatibility of different generations. Constant supply voltage scaling method leads to a high electric field causing short channel effects, mobility degradation, oxide tunneling, and hot carrier degradation. Compromises are currently made to scale at different rates the dimensions of the devices and circuits and the supply voltage. Based on these different theories, the International Technology Roadmap for Semiconductors (ITRS) has set its roadmap for the coming CMOS technology nodes [3].

1.1.2 Potential limitations

Several factors may limit the shrinking of device dimensions. The complexity of integrated circuits (ICs) increases exponentially through the years. This accelerates the difficulties of designing and testing ICs that might become a barrier to the downscaling. Another limiting factor might be the increased overall manufacturing cost of such ICs. This investment is only worthwhile if the revenues are larger than the cost. However, due to the reduction of the supply voltage, the increase in performance is slowing down through the generations, which might level off the manufacturing cost. Finally, physical and electrical effects, such as the short channel effects or the exponential increase of the gate leakage current arising in deep submicron technologies, are increasing the circuit noise. Additional spaces on the chip have to be devoted to reduce this noise and extra components like decoupling capacitances must be added. These effects make therefore the scaling of MOSFET questionable. Moreover, the exponential increase of the gate leakage current will limit the downscaling of MOSFET for mobile applications that have stringent power dissipation requirements.

1.2 Challenges to dielectric scaling

One of the key elements enabling the scaling of the MOSFET is certainly the gate dielectric that isolates the transistor gate from the channel. The gate dielectric plays a fundamental role in the concept "field effect" control (i.e. on the control of short channel effects). The material and electrical properties of the gate dielectric is directly linked to the transistor performance.
Over the past decades, silicon dioxide (SiO_2) has been the natural insulator of the Si-based MOSFETs. The material benefits of SiO_2 are the thermal and mechanical robustness of the film, which can withstand the aggressive environments of device fabrication and the easy processing to grow uniform thin film. The electrical benefits arise from the barrier height of SiO_2 (about 9 eV), which is the result of the difference in bandgaps, being almost ideally distributed between the conduction and the valence bands for compensating the differences in the electron and hole masses. The interface of Si and SiO_2 has excellent properties allowing for high carrier mobility, low interface states, and low trap generation. As illustrated in Table 1.1, which is extracted from the ITRS [3], the next generations of Si-based MOSFETs will require gate dielectrics with thicknesses below 1.5 nm EOT (see definition in Appendix A), both for the high performance logic applications (like microprocessors for personal computers) and low operating power logic applications (like wireless applications).

Technology (nm)	Production year	EOT in nm (see Appendix A)	
		High performance logic	Low operating power logic
130	2002	1.2-1.5	1.8-2.2
107	2003	1.1-1.4	1.6-2.0
90	2004	0.9-1.4	1.4-1.8
80	2005	0.8-1.3	1.2-1.6
70	2006	0.7-1.2	1.1-1.5
65	2007	0.6-1.1	1.0-1.4
50	2010	0.5-0.8	0.8-1.2
25	2016	0.4-0.5	0.6-1.0

Table 1.1: *EOT for the future generations of Si-based MOSFET technologies, including high performance logic applications and low power applications, as from [3].*

The use of ultra-thin oxide in deep submicron MOSFET leads to several issues:

- Direct tunneling current
- B penetration phenomenon in p-MOS transistors
- Polysilicon depletion
- Reliability

For ultra-thin gate oxides (< 3 nm), the probability of an electron to tunnel through the entire oxide becomes significant. The gate leakage current increases exponentially with scaling the gate dielectric thickness: about 1 decade increase in gate leakage current for 0.2 nm decrease in oxide thickness. This huge increase in gate leakage current has a strong impact on power consumption and limits the use of SiO_2 to very specific high performance applications. This aspect will be further detailed in chapter 2 of this thesis.
Another main issue in the scaling of the gate oxide thickness is the intrinsic reliability of such films and how the oxide breakdown is related to the device failure. Recent research has shown that oxide breakdown does not necessarily lead to device or circuit breakdown ([4], [5]). It was shown that the reliability criterion strongly depends on the device functionality (analog/digital) and the application (high performance or low power). It seems that the main showstopper for oxide thickness scaling is the exponential increase of the gate leakage current and not the reliability of the thin gate dielectric.

1.3 Solution to dielectric scaling

A solution to the scaling challenges is to move to a material of higher relative dielectric constant (*K* or ε). By implementing a higher *K* material, the physical film thickness can be increased (see definition of the EOT in Appendix A). A physically thicker dielectric film is desirable because of the reduction in gate leakage current, increased B penetration resistance, and improved long-term reliability. There are a number of candidates being pursued as potential high *K* gate dielectrics, including metal oxides, and metal silicates [6]. However, the integration of these dielectrics into a MOS device poses a number of technological challenges. These challenges include:

- good thermal stability in contact with Si preventing the formation of thick SiO_x interfacial layer or silicide layer;
- a sufficiently large energy band gap to reduce gate leakage current;
- low density of intrinsic defects at the Si/dielectric interface and in the bulk of the material to maximize carrier mobility in the channel and achieved the stringent requirement for ten years reliability;
- compatibility with the existing CMOS fabrication process, like for example high thermal budgets;
- and last but not least the adoption of a candidate in the time frame required by the ITRS (Table 1).

Therefore an approach that builds on the established understanding of the present MOS dielectric materials, namely silicon nitride and silicon dioxide, are required. Silicon nitride (Si_3N_4) is a strong candidate and possesses a number of desirable properties for application in the ultra-thin gate dielectric arena. The relative dielectric constant is 7.5 for Si_3N_4 compared to 3.9 for SiO_2. The Si_3N_4 film is denser than SiO_2, which makes it a better diffusion and implant barrier (e.g. better B penetration resistance). However, Si_3N_4 has the major drawback of having a poor interface with

the underlying Si substrate resulting in high interface state densities. Also, Si_3N_4 films have high bulk traps which give rise to an additional trap assisted leakage mechanism ([7], [8]).
To leverage the benefits of Si_3N_4, and minimize the shortcomings, a hybrid approach can be used, called oxynitride or nitrided oxide ($Si_xO_yN_z$). The ideal oxynitride film would have the Si interface quality of SiO_2, with the increased dielectric constant and boron pentration resistance of Si_3N_4, along with no channel carrier mobility degradation and excellent reliability. The realization of an oxynitride dielectric enables thus an extention of the life of SiO_2 based semiconductor processing.

1.4 Objective and outline of the thesis

The research described in this thesis investigates the scaling limit of nitrided gate oxide thickness and its integration in advanced MOSFET transistors.
The motivations of this work is detailed in chapter 2 where the importance of gate leakage current specifications is underlined. This leakage current specifications pose a stringent requirement on the gate dielectric. It is shown that while pure gate oxide can not meet these gate leakage specifications, plasma nitrided oxides, as an extension of the gate oxide, can fulfill these requirements.
Still, because advanced MOSFETs require ultra-thin plasma nitrided oxides, optimization of electrical (namely Capacitance-Voltage measurements) and physical characterization techniques are required. These two items are detailed in chapter 3 and 4, respectively. These characterization techniques are then used to optimize ultra-thin plasma nitrided oxides, as presented in chapter 5. The impetus for improvement has been to achieve low EOT with low gate leakage current density (J_G) while maintaining high effective carrier mobility. Results on the reliability of these ultra-thin plasma nitrided oxides are summarized. The extendibility of plasma nitrided oxide using new processes is finally discussed. Optimized plasma nitrided oxides have been then integrated in advanced N- and PMOS transistor process flows with a view to maximize short channel transistors performance (chapter 6). In chapter 6, a thorough study of the polysilicon gate activation and deactivation is proposed. Several technological changes are proposed to maximize the gate activation with a view to reduce the total electrical thickness of the gate dielectric. A discussion on the compatibility of the optimized polysilicon gate stack with the formation of ultra-shallow junction in advanced CMOS process flows is then proposed.
Chapter 7 summarizes this thesis work. The future of plasma nitrided oxide with various potential CMOS architectures is finally discussed. Views on the scalability limit of CMOS transistors are proposed.

1.5 References

[1] G.E. Moore, *Lithography and the future of Moore's law*, Proc. SPIE, Vol. 2437, pp. 2-17, 1995.

Chapter 1

[2] R.D. Isaac, ???, IBM J. Res. Dev., Vol. 44, pp. 369-???, 2000.

[3] International Technology Roadmap for Semiconductors, 2001 Edition, Semiconductor Industry Association, San Jose, CA 95129.

[4] B. Kaczer, R. Degraeve, M. Rasras, K. van de Mieroop, P.J. Roussel and G. Groeseneken, *Impact of MOSFET gate oxide breakdown on digital circuit operation and reliability*, IEEE Trans. Electron Devices, Vol. 49, pp. 500-506, 2002.

[5] B. Kaczer and G. Groeseneken, *Potential vulnerability of dynamic CMOS logic to soft gate oxide breakdown*, IEEE Electron Device Lett., Vol. 24, pp. 742-744, 2003.

[6] D.A. Muller, T. Sorsch, S. Moccio, F.H. Baumann, K. Evans-Lutterodt and G. Timp, ??, Nature, Vol. 399, p. 758, 1999.

[7] S.M. Sze, VLSI Technology. 2nd ed. 1988, New York: McGraw-Hill Book Co.

[8] M. Ino, N. Inoue, and M. Yoshimaru, *Silicon nitride thin-film deposition by LPCVD with in situ HF vapor cleaning and its application to stacked DRAM capacitor fabrication*, IEEE Trans. on Electron Dev., Vol. 41, pp. 703 – 708, 1994.

Chapter 2
Impact of the gate leakage current in sub-micron CMOS transistors

2.1 Introduction

2.1.1 Motivation

High leakage current for sub-micron CMOS technologies is becoming a significant contributor to power dissipation of CMOS circuits as V_T, L_G and T_{ox} are reduced. Leakage currents arise due to the incomplete turn-off of transistors in their subthreshold region (i.e. where the gate-source voltage, V_{GS}, is less than the threshold voltage, V_T).

Gate leakage is predicted to increase at a rate of more than 100× per technology generation, while sub-threshold (or transistor) leakage increases by around 5× for each technology generation [1]. This could result in gate leakage becoming the dominant contributor to leakage current.

Furthermore, minimizing power dissipation is becoming very important because of the large complexity of ASICs that contain potentially millions of gates [2]. Low power dissipation/gate is thus required, preventing that the total package dissipation level is not exceeded.

Consequently the identification of the various gate leakage components as well as their contribution to the total power dissipation is very important for estimation and reduction of leakage power, especially for low-power applications (battery-driven equipment such as mobile phones).

This section will start with a brief description of the various leakage current components in a sub-micron MOSFET. The gate leakage current is then thoroughly studied as well as its impact on transistor performance. N-MOSFETs will be studied as they exhibit a worse gate leakage current scenario. Moreover, the impact of the gate leakage current on a standard 6T SRAM cell has been investigated. An estimation of the impact of the gate leakage current on the total standby power

dissipation on a chip is then presented, based on the extrapolation of the 0.12 µm CMOS technology at Philips. Finally, some circuit techniques and technological changes are discussed with the view to reduce the gate leakage current.

2.1.2 Overview of transistor leakages

Leakage currents may come in many forms including transistor leakage, gate leakage, isolation related leakage and junction leakage, all of which are shown schematically in Figure 2.1. A brief definition of the various leakage current components is done below. The gate leakage current component is detailed in the next section.

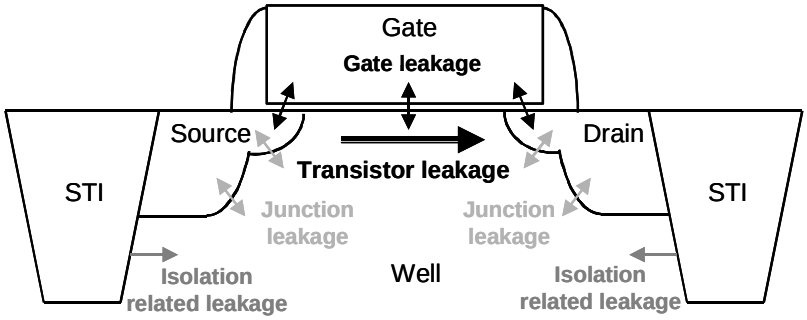


Figure 2.1: Schematic representation of leakage currents in a MOSFET

Advanced short channel MOSFETs require the use of heavily doped shallow junctions in combination with highly doped halos for better short channel effects (SCE). This will create significant band-to-band tunneling (BTBT) current through the drain-well junction. This BTBT leakage has been identified as the main contributor of the pn junction leakage [1].

Transistor leakage (also called source-drain leakage or subthreshold leakage) is the main contributor to the total off-state leakage current (I_{OFF}) although in sub-50 nm transistors, its first place is seriously in competition with the gate and junction leakage currents, as detailed later in this chapter.

When a high drain voltage is applied to a short channel device, it lowers the source potential barrier, resulting in a decrease of V_T, (Figure 2.2). The source then injects carriers into the channel surface independently of the gate voltage and increases exponentially the off-state leakage current. This effect is known as Drain Induced Barrier Lowering (DIBL). As presented in Figure 2.2(b), the DIBL dramatically increases when scaling the channel length of transistors.

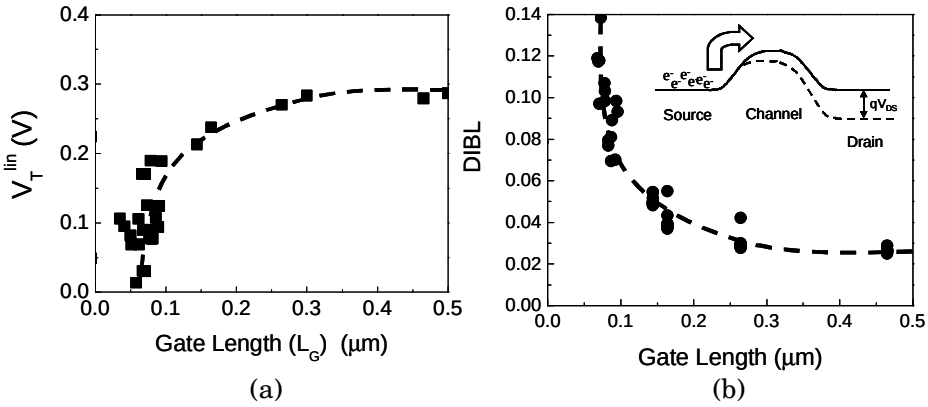


Figure 2.2: Impact of MOSFET channel length on (a) linear V_T (V_{DS}=50 mV) and (b) DIBL. A schematic of the DIBL phenomenon is shown in the inset of Figure (b).

Shallow Trench Isolation is the only lateral isolation scheme that meets the requirements of deep sub-micron technologies in terms of active area scaling and topography at gate level [4]. The abrupt isolation/active area transition induces an enhanced 2-D control of the gate on the corner of the transistor. This can lead to a local decrease in the V_T at the channel edges, resulting in a parasitic leakage path along the lateral transistor ([5] and [6]).
Note that in sub-90 nm CMOS technologies, transistor, junction and gate leakages are dominating the total leakage current. In this work, the impact of the gate leakage current on the total leakage current has been studied.

2.2 Impact of the gate leakage current on transistor characteristics

2.2.1 Definition of the direct tunneling current

The high electric field coupled with ultra-thin gate dielectrics (<3 nm) yields tunneling of electrons (or holes) from the gate to the bulk or to the LDD-to-gate overlap and vice versa. Electrons (or holes) tunnel through a trapezoidal potential barrier. In Figure 2.3(a) and (b), a schematic band diagram is shown for an NMOS in inversion and accumulation, respectively.

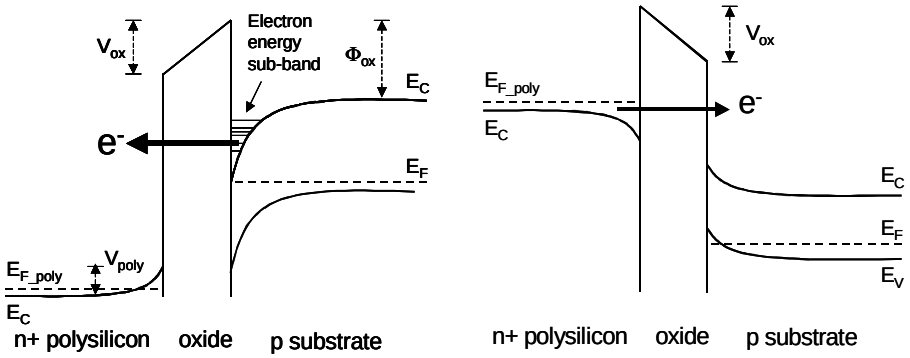


Figure 2.3: *Schematic of the band diagrams of an NMOS structure showing the quantization effects of the substrate electron energy and the direct tunneling of the electrons from (a) the substrate inversion layer to the polysilicon gate (V_{poly} is the voltage drop in the polysilicon layer due to the polysilicon depletion effect) and (b) the substrate accumulation layer to the polysilicon gate.*

For ultra-thin oxide films, the probability for an electron (or a hole) to tunnel through the trapezoidal barrier of the oxide becomes significant ([7], [8]) and increases exponentially with decreasing the oxide thickness (see Equation 2.1).

$$T = \exp\left(-\frac{4}{3}\sqrt{\frac{2m^*}{\hbar^2}}\frac{t_{ox}}{qV_{ox}}\Phi_b^{3/2} \right) \tag{2.1}$$

where T, m^*, t_{ox} and Φ_b are the tunneling probability [7], the effective mass, the oxide thickness and the barrier height respectively. The parameters $\hbar$, q and V_{ox} represent Planck's constant, the electronic charge and the potential drop across the oxide, respectively.

In Figure 2.4(a), the gate leakage current has been measured for various gate oxide thicknesses. Because the gate leakage current is an exponential function of the electric field across the gate oxide (Equation 2.1), the gate leakage current shows an exponential dependence on the gate-to-source (V_{GS}) bias, as shown in Figure 2.4(a). The dependence of the gate current on the drain bias (V_{DS}) for two different oxide thicknesses has been also investigated. For a given gate oxide, increasing the drain bias yields a decrease of the gate current, as shown in Figure 2.4(b). This can be explained by the fact that higher drain voltage reduces the electric field across the gate oxide at the drain end of the channel (lower V_{ox} in Equation 2.1), therefore a lower gate leakage current is measured under high V_{DS}. It can be also observed that there is a strong dependence of V_{DS} in the regime $V_{FB} \leq V_{GS} \leq 0$ V. This effect will be further discussed in the next section.

From these observations, it can be concluded that the gate leakage of a transistor for a given technology (i.e. a given gate dielectric thickness) is determined by V_{GS} and V_{GD} biases seen by the device.

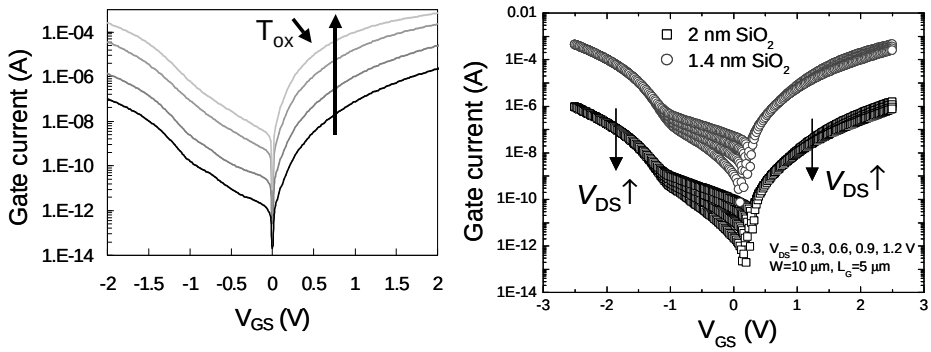
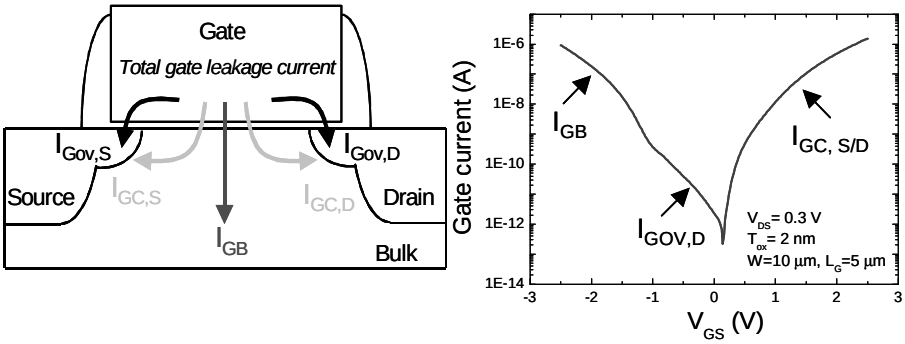


Figure 2.4: *(a) Gate leakage current measured on 10×10 µm² NMOS transistors having various gate oxide thicknesses as a function of the gate-to-source bias (V_{GS}). (b) Gate leakage current measured as a function of V_{GS} at various V_{DS} on NMOS transistors with a 2 or 1.4 nm gate oxide. The channel length and width of the transistors are 5 and 10 µm, respectively. For both figures, the bulk and source biases are set to 0 V.*

2.2.2 Components of the gate direct tunneling current

The gate direct tunneling current can be divided into three major components (Figure 2.5(a)): the gate-to-channel (I_{GC}) flowing through the source/drain extension area, the gate-to-bulk (I_{GB}) and the gate-to-source/drain extension overlap regions ($I_{GOV,\ S/D}$) currents. In Figure 2.5(b), the gate leakage has been measured as a function of V_{GS}. Three main regimes can be observed: $V_{GS}<V_{FB}$, $V_{FB}\leq V_{GS}\leq 0$ V, $V_{GS}>0$. The resulting three gate leakage components have been added according to MM11 modeling results as published in [9]. It was shown that the tunneling gate current for $V_{GS}\leq 0$ V consists of $I_{GOV,S/D}$ and I_{GB}. One can observed a kink around $V_{GS}\approx V_{FB}$, which indicates the transition of dominance from $I_{GOV,S/D}$ and I_{GB}. This is also confirmed in Figure 2.4(b), where the impact of V_{DS} on the gate leakage current is more pronounced for $V_{FB}\leq V_{GS}\leq 0$ V. The gate leakage current component at 0 V is $I_{GOV,S/D}$. For $V_{GS}>0$ V, I_{GC} is dominating. The respective impact of the gate leakage components on the total gate leakage current as a function of the transistor regime is summarized in Table 2.1.



(a) (b)

Figure 2.5: *(a) A schematic showing the various gate leakage components of a MOSFET. The gate-to-source/drain overlap (I_{Gov}), the gate-to-channel (I_{GC}) and gate-to-bulk (I_{GB}) currents are represented. (b) Gate leakage current components as a function of V_{GS} for NMOS transistor having a 2 nm gate oxide. The channel length and width are 5 and10 µm, respectively. The gate leakage components have been added using MM11 [9].*

	Accumulation $V_{GS}<V_{FB}$ ($V_{GS}>V_{FB}$)	Weak-inversion or depletion $V_{FB}<V_{GS}<V_T$ ($V_{FB}>V_{GS}>V_T$)	Strong inversion $V_{GS}>V_T$ ($V_{GS}<V_T$)
I_{GB}	electrons (electrons)	-	-
I_{GC}	-	-	electrons (holes)
$I_{GOV,\ S/D}$	electrons (holes)	electrons (holes)	electrons (holes)

Table 2.1 *The role of the different gate leakage current components as a function of the MOSFET operation regimes. The type of carriers involved in each gate tunneling component is mentioned for NMOST and in between brackets for PMOST.*

2.3 Impact of gate tunneling current on the transistor off-state behavior

I_{OFF} is the drain current when the gate, bulk and source voltages are zero and the drain bias is equal to V_{DD}. I_{OFF} is influenced by the threshold voltage, the gate length, the channel/surface doping profile, the source/drain junction depth, the gate oxide thickness and by V_{DD}.

2.3.1 Definition of the gate leakage current component in the off-state regime

In the case of NMOS transistors in the off-state regime, both the n+ polysilicon gate and the n+ source/drain regions are highly doped and become degenerate, resulting in a n+/insulator/n+ MOS tunneling structures between the gate and the source/drain extension region. Its band diagram is shown in Figure 2.6(a). The $I_{GOV,\ S/D}$ has been measured for NMOS transistors with two gate dielectric thicknesses. It can be observed in Figure 2.6(b) that $I_{GOV,\ S/D}$ increases dramatically with scaling the gate dielectric thickness. Moreover, $I_{GOV,\ S/D}$ is constant over the channel lengths range. This indicates that, in this particular case, the source/drain extension overlap is constant over the whole channel lengths range. This independency of $I_{GOV,\ S/D}$ with the channel length is no more valid if the dose of the halos is strong. Indeed, when scaling the transistor channel length, highly doped halos can overlap inducing a variation of the source/drain extension overlap. This will result in a decrease of $I_{GOV,\ S/D}$ for short channel devices. This variation is however very small compared to the large changes observed in $I_{GOV,\ S/D}$ when varying the gate dielectric thickness. A variation in the measurement of short channel lengths could also induce small variations in $I_{GOV,\ S/D}$.
Because $I_{GOV,\ S/D}$ does not strongly depend on the transistor gate length, it can be approximated that $I_{GOV,\ S/D}$ is only a function of the gate leakage current density (J_G) related to the gate dielectric and of the source/drain extension overlap length (X_{OV}), as presented in Equation 2.2:

$$I_{GOV,S/D} = X_{ov} \times J_G \tag{2.2}$$

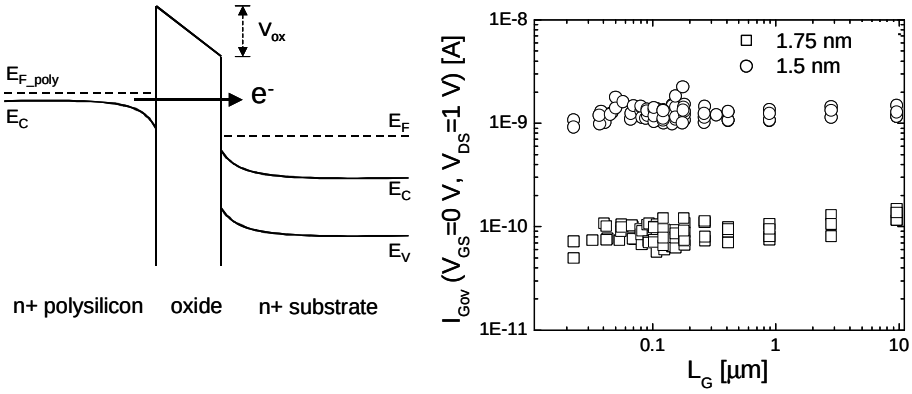


Figure 2.6: *(a) A schematic of the band diagram of a n+ polysilicon/SiO_2/n+ Si MOS structure showing electron tunneling from the gate to the source/drain extension (V_{FB} ≈0 V). (b) Overlap gate current measured on NMOS transistors having various gate lengths and gate dielectric thickness (1.5 and 1.75 nm) but same source/drain overlap areas. Several NMOS transistors have been measured over the wafer for each gate length. The width of the transistors is fixed at 10 µm.*

2.3.2 Effects of the gate leakage current on the transistor off-state behavior

I_{OFF} has been measured at various gate lengths for NMOS with plasma nitrided oxides (Figure 2.7(a)). For long channel devices, I_{OFF} is strongly dependent on the gate dielectric thickness and is constant over the gate lengths. For short channel transistors, I_{OFF} is increasing exponentially as the result of uncontrolled SCE (related to the V_T roll-off phenomenon, see section 2.1.2). Note that the spread seen for the smallest gate nitrided oxide (EOT~1.1 nm) can be attributed to the non uniformity of this dielectric over the wafer.

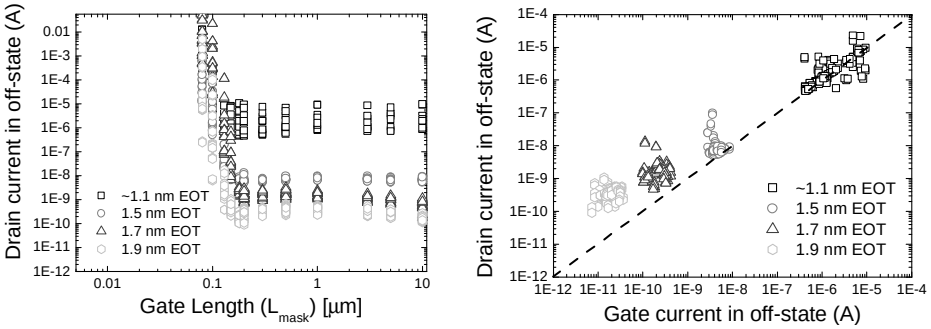


Figure 2.7: *(a) Drain current in the off-state regime (V_{GS}=V_{SS}=V_{BS}=0 V, V_{DS}=V_{DD}=1 V) versus gate length (mask length) for NMOS transistors with various SiON gate dielectric thicknesses ranging from 1.9 down to 1.1 nm EOT (see Appendix A). Several transistors have been measured for each gate length. (b) Drain current versus gate current in the off-state regime for long channel NMOS transistors ($L_{G,mask}$>0.13 µm).*

The impact of the gate leakage current on the total I_{OFF} has been investigated for the NMOS transistors with various gate dielectric thicknesses. For this purpose, only long channel transistors ($L_{G,mask}$>0.13 µm) have been selected in order to avoid variations of I_{OFF} due to uncontrolled SCE. In Figure 2.7(b), I_{OFF} is plotted as a function of the off-state gate leakage current (I_{G_OFF}). For NMOS devices with a gate dielectric thicker than 1.5 nm EOT, I_{OFF} is larger than I_{G_OFF}. The gate leakage current starts to dominate the total I_{OFF} for plasma nitrided oxides below 1.5 nm EOT (Figure 2.7(b)). As already mentioned in the previous section, the gate leakage current component in the off-state regime (at V_{GS}=0 V) is $I_{GOV,\ S/D}$ that flows in the gate-to-drain overlap region.

As a consequence, for transistors having an ultra-thin gate dielectric, I_{OFF} is no more equal to the transistor leakage ($I_{off}^{S,D}$) but I_{GOV} should be added as a second important component:

$$I_{off} = I_{off}^{S,D} + I_{GOV,S/D} \tag{2.3}$$

where I_{off} is the total off-state current, $I_{off}^{S,D}$ is the transistor leakage and $I_{GOV,\ S/D}$ is the overlap gate leakage current. I_{OFF} is fixed for a given circuit application (e.g. high performance or low power).

From Equation 2.2 and considering that X_{ov} is about 15 % of the total gate length of the transistor [1], the impact of the relative contribution of $I_{GOV,\ S/D}$ on the drive current of the transistor at fixed I_{OFF} can be estimated. In Figure 2.8(a), this impact has been calculated, using calibrated analytical model, for transistors having a gate dielectric with an EOT of 1.2 nm with a J_G of 17 A/cm^2 (at V_{GS}=1 V), a transistor gate length of 40 nm and with an I_{OFF} specification at 10^{-9} A/µm, as indicated in [1] for the 65 nm CMOS technology node. The resulted I_{ON} has been compared to a

device with non-leaky gate dielectric. In Figure 2.8(a), the reduction in I_{ON} as a function of the contribution of $I_{GOV,\ S/D}$ to I_{OFF} is shown. An increase of the contribution of $I_{GOV,\ S/D}$ to the total off-state leakage current forces a decrease of the transistor leakage, $I_{off}^{S,D}$, given that the total off-state current is fixed. A decrease of $I_{off}^{S,D}$ results in a reduction of V_T since the supply voltage is also fixed for each circuit application. The decrease of V_T results in a dramatic decrease in I_{ON}.
As an example, in order to limit the impact of the gate leakage current on the drive current of short channel transistors to 5 % reduction, $I_{GOV,\ S/D}$ should be limited to less than 30 % of the total I_{OFF}. Note that similar results have been obtained for other devices (other type of applications) described in [1] for the 65 nm node technology.

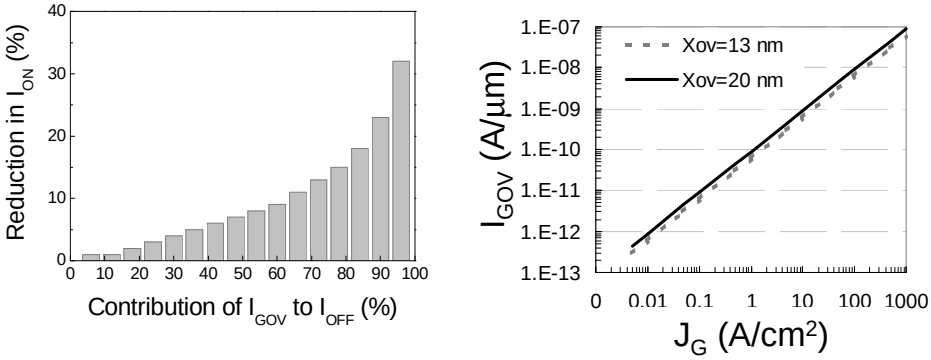


Figure 2.8: *(a) Impact of the relative contribution of I_{GOV} to fixed I_{OFF} on I_{ON} reduction (in comparison with a non leaky gate dielectric). I_{GOV} =15%×L_G×J_G. (b) Impact of the gate current density (J_G) on $I_{GOV,S/D}$ for NMOS devices with an overlap length of 13 or 20 nm.*

The impact of the gate leakage current density (J_G) on $I_{GOV,\ S/D}$ has been then studied for transistors having a gate-to-drain/source overlap of 13 or 20 nm, as presented in Figure 2.8(b). As expected from Equation 2.2, $I_{GOV,S/D}$ has a linear dependency on J_G. From Figure 2.8(a) and (b), requirements of J_G at a given I_{OFF} can be estimated. Indeed, as an example, if the I_{OFF} specification is set to 1 nA/µm, J_G should be less than 3 A/cm^2 to maintain the $I_{GOV,\ S/D}$ contribution of less than 30 % to the total I_{OFF} (i.e. to have less than 5 % degradation in I_{ON}). We will see later in this chapter that such EOT-J_G requirement is impossible to meet with pure oxide gate dielectrics. Other gate dielectric materials need therefore to be implemented to reduce J_G. The I_{OFF} specification could be also increased, however this is not always possible for low power applications such as portable applications that require very low I_{OFF}. There is therefore a trade-off between, high performance (i.e. I_{ON}) and limited I_{OFF}.

2.4 Consequences of excessive leakage current on circuit

2.4.1 Impact of the gate leakage current on a 6T SRAM cell leakage

In the previous section, it was shown that the gate tunneling current flowing from the gate through the source/drain extension overlap plays a significant part in the off-state drain current. In this section, the impact of the gate leakage current on the total leakage of a typical six-transistor (6T) cell used for CMOS static random-access memories (SRAM) is studied.

The 6T SRAM cell consists of two cross-coupled CMOS inverters (T1 to T4) that store one bit of information (i.e. function as a latch), and two n-type transistors (T5 and T6) that connect the cell to the bit lines (i.e. provide access to the latch), as presented in Figure 2.9.

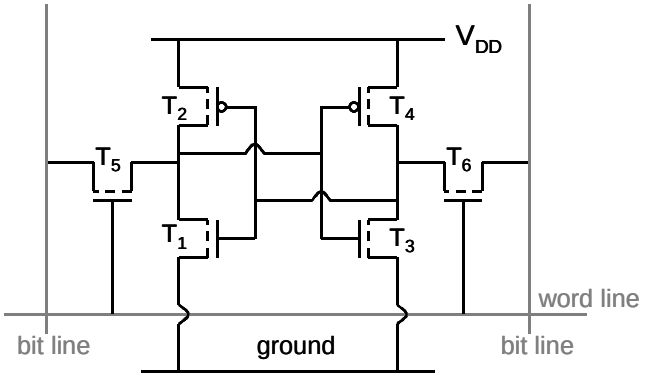


Figure 2.9: *A schematic of a 6T SRAM cell.*

The 6T SRAM cell leakage can be calculated:

$$I_{leak\_6T\,SRAM} = (2 \times W_{pass} \times X_{ov} \times J_G) + (W_{load} \times L_G \times J_G) + (W_{driver} \times I_{OFF} + W_{driver} \times X_{ov} \times J_G)$$
$$+ (W_{load} \times I_{OFF} + W_{load} \times X_{ov} \times J_G) + (W_{driver} \times L_G \times J_G) + (W_{pass}.I_{OFF} + W_{pass} \times X_{ov} \times J_G) \quad (2.4)$$

where $I_{leak\_6T\ SRAM}$ is the leakage current in a 6T SRAM cell. W_{pass}, W_{driver} and W_{load} are the width of the transistors T5 /T6, T1/T2 and T3/T4, respectively. X_{OV}, L_G and J_G are the gate-to-source/drain overlap length, the gate length and the gate current density, respectively. I_{OFF} is the transistor leakage and gate-to-drain overlap leakage (see Equation 2.3).

The impact of J_G on the total SRAM cell leakage has been estimated for the CMOS 65 nm technology node. Note that for this purpose, the dimensions of the various

transistor parameters as well as the design of the 6T SRAM cell were estimated by extrapolating typical design and library characteristics of the Philips 0.12 µm technology node. The $I_{leak\_6T\ SRAM}$ has been calculated for various gate length and I_{OFF} specifications. The obtained $I_{leak\_6T\ SRAM}$ have been plotted as function of J_G, as presented in Figure 2.10. For each various L_G and I_{OFF} specification, $I_{leak\_6T\ SRAM}$ is independent of J_G until a certain value for J_G. Above this value, $I_{leak\_6T\ SRAM}$ increases linearly with J_G independently of the channel length and I_{OFF} specifications. This can be explained by the fact that above a certain J_G value, the total leakage current of the 6T cell is dominating by J_G.

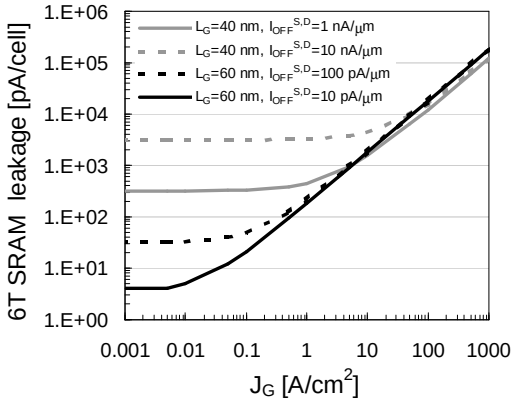


Figure 2.10: *Impact of J_G on the total leakage of a standard 6T SRAM cell for devices targeting the 65 nm CMOS technology node [1].*

2.4.2 Impact of the gate leakage current on the total power dissipation of a chip

Increasing the number of transistors on a chip, while decreasing their size, has been a key factor in developing faster and faster chips over the decades. However, the decrease of the transistors size results in an exponential increase of I_{OFF} both due to the transistor leakage and gate leakage, as seen in the previous section. As a result, the total power consumption of the chip is increasing dramatically. What is important to point out at this stage of the discussion is that the allowed power consumption is very application dependent [1].

In this section, only portable applications will be studied. For these applications, low power dissipation is required. In order to study the power dissipation in a chip using advanced CMOS technologies, an extrapolation of the typical library and design characteristics of Philips for the CMOS 0.12 µm technology node has been done.

For a CMOS circuit, the total power dissipation includes the active and standby power dissipation (Equation 2.6). The active power has two components: a dynamic and a static power (Equation 2.7). The dynamic power is the switching power due to charging and discharging of load capacitance. The static component is the so-called short-circuit power dissipation and is determined by the leakage current through each transistor. In the standby mode, the power dissipation is due to the standby leakage current (Equation 2.8).

$$P_{total} = P_{active} + P_{standby} \tag{2.6}$$

where

$$P_{active} = \frac{1}{2} \times f \times C_{load} \times V_{DD}^2 + \left(I_{mean} \times V_{DD}\right) \tag{2.7}$$

$$P_{standby} = I_{OFF} \times V_{DD} \tag{2.8}$$

and P_{total}, P_{active} and $P_{standby}$ are the total, the active and the standby power dissipation, respectively. C_{load}, α, f and V_{DD} are the load capacitance, the switching activity, the operation frequency and the supply voltage, respectively. I_{mean} is the average short-circuit current. I_{OFF} is the cumulative leakage current due to all the components of the leakage current.

An estimation of the total leakage current in a chip (to be used in mobile applications) for CMOS technologies ranging from the 0.12 μm down to the 45 nm nodes has been made. In Figure 2.11, the active and standby leakage currents are plotted as a function of the different CMOS technology nodes. The exponential increase of J_G, when scaling the technology node, has a dramatic impact on the total leakage current of the chip in the standby mode. We have seen in this chapter that the transistor leakage but also the gate leakage and junction leakage are increasing with scaling CMOS. As we have seen for a 6T SRAM cell in the previous section, above a certain J_G, I_{OFF} is dominated by the gate leakage current. As a consequence, the standby power increases exponentially with the gate leakage current above a certain J_G (see Equation 2.8). The active power is also increasing when scaling the CMOS technology nodes. This increase can be explained by the increase of C_{load} and f with downscaling CMOS technologies. Although the active power depends strongly on V_{DD}, this latter one does not scale enough to compensate for the increase in C_{load} and f. The ratio of the standby power over the active power has been also plotted in Figure 2.11. This ratio is increasing dramatically as a result of the exponential increase of J_G. It is therefore necessary to define a specification for J_G to avoid the dramatic increase of the power consumption. Depending on the application and therefore on the maximum allowed power consumption in the circuit, the J_G requirement will differ.

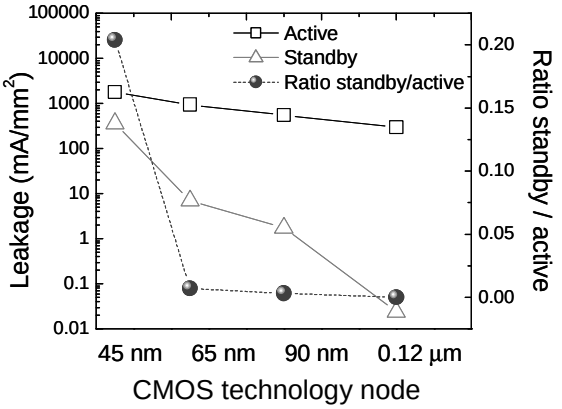


Figure 2.11: *Active and standby leakages trend as a function of CMOS technology nodes for a low power application. The ratio of the two powers is plotted on the second y axis.*

It is now interesting to look at the power consumption due to the gate leakage current. The power consumption due to the gate leakage can be calculated as follows:

$$P_{leak} = J_G \times V_{DD} \times A \tag{2.9}$$

where P_{leak} is the power consumption due to gate leakage current, J_G is the gate current density and A is the electrically active area estimated from [1].

As can be observed in Figure 2.12, the power consumption due to gate leakage current increases exponentially with increasing J_G. Scaling the supply voltage does not help sufficiently in the reduction of P_{leak} and cannot be scaled too much due to the increase of the transistor leakage and the performance of the device. Note that this limitation is very application dependent.

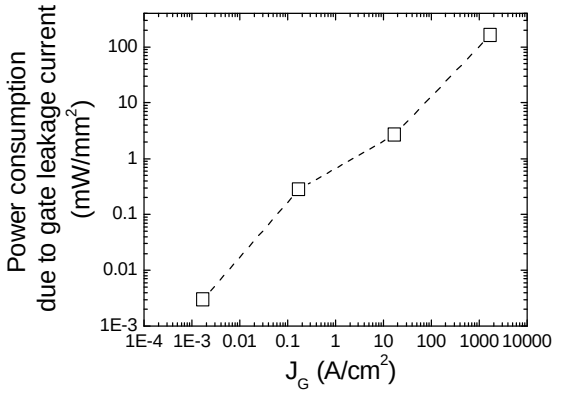


Figure 2.12: *Power consumption estimation due to gate leakage current as a*

It is thus of first importance to realize that for future CMOS technologies, reducing the power supply voltage does not completely solve the issue regarding the total power consumption. There is a trade-off between the performance, the power consumption and the application.

2.5 Leakage reduction techniques

2.5.1 Leakage current reduction by circuit techniques

Low Power design has become a major concern of all ASIC and IC designers in recent years because of the battery lifetime, particularly. Circuit design solutions have been proposed to the high power dissipation issue. Only three out of the numerous solutions will be listed here, others can be found in [10]:

1. switching parts of the system off when not in use, using lower power technologies and architectures.
2. using multiple V_{Ts}: high- and low-V_T transistors are included on the same chip. The high V_T will suppress the transistor leakage while the low V_T transistors will achieve high performance. These multiple V_{Ts} transistors can be achieved by varying the implant conditions of the steep retrograde channel.
3. using multiple oxides: Note that the change in the gate dielectric thickness should be coupled to a change in gate length of the transistor to prevent any severe SCE.

2.5.2 Leakage current reduction by technology changes

For further downscaling of CMOS devices targeting low power applications, solutions in the process can be found to limit the leakage current without drastically changing the CMOS architecture. The solutions proposed here to reduce the off-state leakage seem to be the simplest to implement in a conventional “standard” CMOS process flow.

An obvious way to reduce the gate leakage current is to engineer the gate dielectric. This solution will be detailed in the next section.
The use of metal gates would eliminate polysilicon depletion effects and therefore decrease the electrical thickness of the gate dielectric. For a given electrical thickness, thicker gate dielectric could be used resulting in lower gate leakage current. This solution will be further detailed in chapter 6. Furthermore, the increase in carrier mobility (that is to say in I_{ON}) obtained with, for example, strained Si would enable a thicker electrical thickness to achieve similar performance as in conventional bulk Si substrate using thinner gate dielectric.

While this chapter is focusing on gate leakage current contribution to the total off-state leakage, transistor leakage as well as junction leakage are also important components that need to be reduced.
One way to reduce the transistor leakage is to increase the V_T. As already mentioned in the first part of this chapter, this cannot be achieved by increasing the channel doping of a standard bulk Si as it will dramatically degrade the carrier mobility in the channel and increase the tunneling current between the source and drain p-n junctions with the substrate. Building the transistors on a thin silicon layer on top of an embedded layer of insulation (silicon on insulator, SOI, technology) could be a solution to lower the power dissipation [17]. Indeed, the thin insulator layer reduces the current leakage to a minimum; while the thin silicon substrate can be fully depleted yielding maximum drive current and enabling the transistor to switch on and off faster [18], [19]. Because the advantages of SOI are related to speed gain and reduced power dissipation, SOI is expected to gain more importance in the future when ultra-large scale integration (ULSI) of miniaturized bulk CMOS becomes more and more difficult [20], [21].
Having a more graded profile for the deep junctions will reduce the junction leakage. In the case of NMOS devices, the junctions are very often formed with Arsenic (As) dopants. The profile of such junctions is rather abrupt and will therefore yield high junction leakage. Adding a co-implantation of a lighter element such as Phosphorous (P) will result in a more graded profile and therefore will decrease the junction leakage. Defects in the junctions are also enhancing junction leakage and should be minimized.

As mentioned above, the list of the possible technological changes to lower I_{OFF} that is described here is not exhaustive but can be applied easily in a standard CMOS process flow.

2.5.3 Leakage current reduction by gate dielectric engineering

The obvious solution to reduce the gate leakage current is to increase the dielectric constant (*K*) of the material. This will result in a physically thicker film while still having a thin EOT (see Appendix A). Therefore, the gate leakage current will drop significantly. In the literature, High-*K* dielectrics such as HfO_2 exhibiting a decrease in gate leakage current of 4 orders of magnitude as compared to pure SiO_2 have been reported. However, issues with these dielectrics include the “pinning” of the Fermi-level when used in combination with polysilicon electrodes, high fixed (bulk) charge, high interface state density and the threshold voltage instability due to transient charge trapping. Furthermore, it has been suggested that high-k gate dielectrics result in lower effective mobility of inversion charges in Si-based MOSFETs because of soft phonon scattering ([12], [13], [14] and [15]).

Also, one important consideration regarding the integration of such high-*K* material in a circuit is its behavior at high frequencies. Indeed, the *K* value of a material is measured at low frequency and it is not sure that the same *K* value will be obtained at high frequencies. If the high *K* value cannot be obtained at high frequencies, the capacitance and therefore the performance of the transistor will be reduced [16].
As a consequence, nitrided oxides are preferred over higher-*K* gate dielectrics as a solution to decrease the gate leakage current. Various nitridation processes have been studied yielding different nitrogen (N) concentration in the gate dielectric layer ([24], [25] and [26]). Plasma nitridation has been reported as a way to introduce a high amount of N in the dielectric [22]. A higher N content will result in an increase of the dielectric constant. Due to the higher *K* of the nitrided oxide, physically thicker layer for a given EOT, yielding a decrease in gate leakage current. A benchmark of the J_G-EOT trend obtained for NMOS devices with a pure oxide, a furnace nitrided oxide or a plasma nitrided oxide is shown in Figure 2.13. It can be observed that for a given EOT, a decrease of J_G of about a factor 10 is observed when replacing a pure oxide gate dielectric with a plasma nitrided one. Because plasma nitrided films have a higher amount of N as compared to furnace nitrided oxides, a lower J_G is measured for NMOS devices with a plasma nitrided oxide. However, it has been reported that the nitridation of silicon dioxide induces positive fixed charges that yield a shift in a shift of $\mathbf{V}_T$ ([27], [28]). Moreover, it reduces both electron and hole effective mobility if some nitrogen atoms are located at the dielectric/substrate interface [29].

As a conclusion, plasma nitrided oxide could be used as an intermediate medium *K* dielectric for leakage current reduction providing that the N concentration profile is not localized at the gate dielectric/Si interface which yield carrier mobility degradation. Optimization of theses nitrided oxides will be presented in chapter 5.

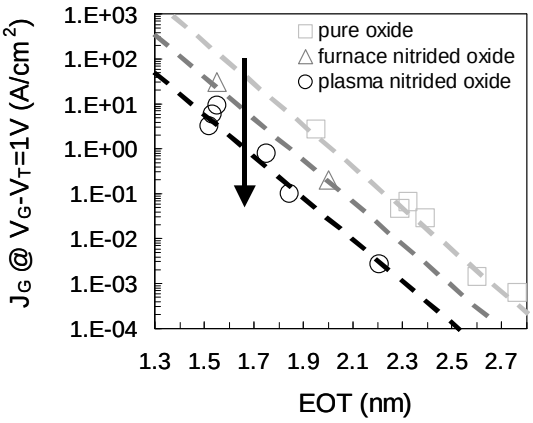


Figure 2.13: *(a) J_G-EOT trend for NMOS devices with a pure oxide, a furnace or a plasma nitrided oxide.*

2.6 Conclusions

The exponential increase of the gate leakage current when scaling the gate dielectric thickness is playing a major role in short channel CMOS transistors. The off-state current is no more equal to the transistor (or subthreshold) leakage only but also includes the gate-to-source/drain overlap leakage and junction leakage (mainly BBT leakage). The impact of the gate leakage current for low power applications of advanced CMOS technology nodes (sub-0.12 μm) has been investigated. It was shown that the gate-to-source/drain overlap leakage current component should be less than 30 % of the total off-state current in order to prevent serious performance degradation (on-state current degradation < 5 %).
An example of the strong dependence of the gate leakage current on the total leakage of a standard 6T SRAM cell has been presented for low power applications [1]. It was shown that there is a value for J_G where the total cell leakage is dominated by the gate leakage. As a consequence the power consumption in the standby mode due to gate leakage current increases exponentially above this J_G value. The maximum J_G allowed in a chip need therefore to be specified for each CMOS application in order to prevent a dramatic increase of the power consumption. When scaling CMOS technology nodes, changes in the design and in the process have to be implemented to reduce the leakage as reducing the supply voltage is no more sufficient. There is a trade-off between the performance, the power consumption and the application.
Introducing a gate dielectric with a higher dielectric constant will reduce the gate leakage current component. It was shown that gate leakage current can be reduced by a factor 10 when using plasma nitrided oxide relative to pure oxide gate dielectric. Plasma nitrided oxide has been chosen as a good intermediate "medium *K*" solution before the introduction of higher *K* material. This nitridation has been selected for this thesis work to form ultra-thin nitrided oxides. The optimization of such thin gate dielectrics is detailed in chapter 5.

2.7 References

[1] International Technology Roadmap for Semiconductors, 2001 Edition, Semiconductor Industry Association, San Jose, CA 95129.

[2] G.E. Moore, *Cramming more components onto integrated circuits*, Electronics, Vol. 38, No. 8, pp. 114-117, 1965.

[3] Y. Taur and T.H. Ning, *Fundamentals of modern VLSI devices*, New-York: Cambridge Univ. Press, Ch. 3, p. 130, 1998.

[4] A. Perera and J.H. Lin, *Trench isolation for 0.45 μm active pitch and below*, IEDM Tech. Digest, pp. 679-682, 1995.

[5] N.Shigyo, S. Fukuda, T. Wada, K. Hieda, T. Hamamoto, H. Watanabe, K. Sunouchi and H. Tango, *Three-dimensional analysis of subthreshold swing and transconductance for fully recessed oxide (trench) isolated ¼-μm-width MOSFETs*, IEEE Trans. Electron Devices, Vol. 35, pp. 945-951, 1988.

[6] P. Sallagoity, M. Ada-Hanifi, M. Paoli and M. Haond, *Analysis of width edge effects in advanced isolation schemes for deep submicron technologies*, IEEE Trans. Electron Devices, Vol 43, pp. 1900-1906, 1996.

[7] S. Nagano, M. Tsukiji, K. Ando, E. Hasegawa and A. Ishitani, *Mechanism of leakage current through nanoscale SiO2 layer*, Journal of Applied Physics, Vol. 75, pp. 3530-3535, 1994.

[8] M. Depas, B. Vermeire, P.W. Mertens, R.L.V. Meirhaeghe and M.M. Heyns, *Determination of tunneling parameters in ultra-thin oxide layer poly Si/SiO2/Si structures*, Solid State Electronics, Vol. 38, pp. 1465-1472, 1995.

[9] R. van Langevelde, A.J. Scholten, R. Duffy, F.N. Cubaynes, M.J. Knitel and D.B.M. Klaassen, *Gate current: modeling, ΔL extraction and impact on RF performance*, IEDM Tech. Dig., pp. 289-292, 2001.

[10] H. Veendrick, *Deep Submicron CMOS ICs: from Basics to ASICs*, Kluwer, Deventer, The Netherlands, 2nd ed., 1998.

[11] V. De and S. Borkar, *Technology and design challenges for low power and high performance*, Proc. Int. Synp. Low Power Electronics and Design, pp. 163-168, 1999.

[12] M. Copel, M. Gribelyuk and E. Gusev, *Structure and stability of ultrathin zirconium oxide layers on Si(100)*, Appl. Phys. Lett., Vol. 76, pp. 436-439, 2000.

[13] M. Gutowski, J.E. Jaffe, C.L. Liu, M. Stoker, R.I. Hegde, R.S. Rai and P.J. Tobin, *Thermodynamic stability of high-K dielectric metal oxides*

ZrO_2 and HfO_2 in contact with Si and SiO_2, Appl. Phys. Lett, Vol. 80, pp. 1897-1890, 2002.

[14] J. H. Lee, Y.S. Kim, H.S. Jung, J.H. Lee, N.I. Lee, H.K. Kang, J.H. Ku, H. Kang, Y.K. Kim, K.H. Cho and K.P. Suh, *Poly-Si Gate CMOSFETs with HfO_2-Al_2O_3 Laminate Gate Dielectric for Low PowerApplications,* VLSI Tech. Digest, pp. 84-85, 2002.

[15] S. Pidin Y. Morisaki, Y. Sugita, T. Aoyama, K. Irino, T. Nakamura and T.Sugii, *Low Standby Power CMOS with HfO_2 Gate Oxide for 100-nm Generation, VLSI Tech. Digest*, pp. 28-29, 2002.

[16] B. Doyle, R. Arghavani, D. Barlage, S. Datta, M. Doczy, J. Kavalieros, A. Murthy and R. Chau, *Transistor elements for 30 nm physical gate lengths and beyond*, Intel Technology Journal, Vol. 6, pp. 42-54, 2002.

[17] J.P. Colinge, S*ilicon-On-Insulator technology: Materials to VLSI*, Kluwer, Boston, Academic Publ., 2nd ed., 1997.

[18] A.G. Aipperspach, D.H. Allen, D.T. Cox, N.V. Phan and S.N. Storino, *A 0.2-µm, 1.8-V, SOI, 550-MHz, 64-b PowerPC microprocessor with copper interconnects*, IEEE J. Solid-State Circuits, Vol. 34, No. 11, pp. 1430-1435, 1999.

[19] S.B. Park, Y.W. Kim, Y.G. Ko, K.I. Kim, I.K. Kim, H.-S. Kang, J.O. Yu and K.P. Suh, *A 0.25 µm, 600-MHz, 1.5-V, fully depleted SOI CMOS 64-bit microprocessor*, IEEE J. Solid-State Circuits, Vol. 34, No. 11, pp. 1436-1445, 1999.

[20] K. Shimomura, H. Shimano, N. Sakashita, F. Okuda, T. Oashi, Y. Yamaguchi, T. Eimori, M. Inuishi, K. Arimoto, S. Maegawa, Y. Inoue, S. Komori and K. Kyuma, *A 1-V 46-ns 16-Mb SOI-DRAM with body control technique*, IEEE J. Solid-State Circuits, Vol. 32, No. 11, pp. 1712-1720, 1997.

[21] N. Shibata, M. Watanabe, Y. Sato, T. Ishihara and Y. Komine, *A 2-V 300-MHz 1-Mb current-sensed double-density SRAM for low-power 0.3-µm CMOS/SIMOX ASICs*, IEEE J. Solid-State Circuits, Vol. 36, No. 10, pp. 1524-1537, 2001.

[22] F.N. Cubaynes, C.J.J. Dachs, C. Detcheverry, A. Zegers, V.C. Venezia, J. Schmitz, P.A. Stolk, M. Jurczak, K. Henson, R. Degraeve, A. Rothschild, T. Conard, J. Petry, M. Da Rold, M. Schaekers, G. Badene, L. Date, D. Pique, H.N. Al-Shareef and R.W. Murto, *Gate dielectrics for high performance and low power CMOS SoC applications*, Proceedings of the ESSDERC Conf., pp. 427-430, 2002.

[23] M.A. Alam, J. Bude and A. Ghetti, *Field accelaration for oxide breakdown - Can an accurate anode hole injection model resolve the E vs. 1/E controversy?*, Proceedings of the IRPS Conf., pp. 21-26, 2000.

[24] E.P. Gusev, H.C. Lu, E. Garfunkel, T. Gustafsson and M. Green, *Growth and characterization of ultrathin nitrided silicon oxide films*, IBM J. Res. Dev., Vol. 43, pp.265-286, 1999.

[25] F.H.P.M. Habraken and A.E.T. Kuiper, ???, Mat. Sci. eng., Vol. R12, pp. 123-??, 1994.

[26] M.M. Moslehi and K.C. saraswat, ???, IEEE Trans. Electron Devices, Vol. 32, pp. 106-??, 1985.

[27] C.T. Chen, F.C. Tseng, C.Y. Chang and M.K. Lee, *Study of electrical characteristics on thermally nitrided SiO2 (nitroxide) films*, J. Electrochem. Soc., Vol. 131, pp. 875-877, 1984.

[28] M.A. Schmidt, F.L. terry, B.P. Mathur and S.D. Senturia, *Inversion layer mobility of MOSFET's with nitrided oxide gate dielectrics*, IEEE Trans. On Elet. Dev., Vol. 35, pp. 1627-1632, 1988.

[29] S. C. Song, H. F. Luan, Y. Y. Chen, M. Gardner, J. Fulford, M. Allen and D. L. Kwong, *Ultra Thin (<20Å) CVD Si3N4 Gate Dielectric for Deep-Sub-Micron CMOS Devices*, IEDM tech. Dig., pp.373-376, 1998.

Chapter 3
Capacitance-Voltage measurements under high gate leakage current

3.1 Introduction

3.1.1 Motivation

It was shown in the previous chapter, that gate leakage current increases exponentially as the gate oxide is scaled. The replacement of pure gate oxides by highly nitrided oxides reduces the gate leakage current by a factor 10. Yet, when scaling down the gate dielectric, the gate leakage current is increasing exponentially exceeding 10 A/cm^2 (at V_{GS}=1 V) for devices with plasma nitrided gate dielectrics below 1.3 nm EOT.

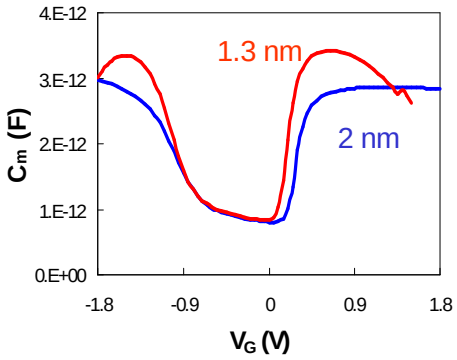


Figure 3.1: *Measured capacitance (C_m) as a function of the gate voltage for NMOS devices with a 2 or 1.3 nm plasma nitrided oxide as the gate dielectric.*

Capacitance-voltage (C-V) characteristics have been measured on NMOS transistors having a 2 or 1.3 nm EOT plasma nitrided oxide, as shown in Figure 3.1. For the NMOS device with thinnest plasma nitrided oxide, and therefore the leakiest, capacitance attenuation in both inversion and accumulation are observed as result of the large direct tunneling current. C-V characteristic is therefore strongly affected by the high gate leakage current. The measured capacitance (C_m) changes due to high gate leakage current while the real intrinsic capacitance is independent of the leakage current.

C-V measurements on MOS structures are essential in the process of research and development of new CMOS generations, as they allow determining characteristic parameters of the gate dielectric such as its thickness (Figure 3.2(a)), fixed charge density, and its interface state density. Furthermore, the active dopant concentration in the polysilicon gate and in the silicon substrate can be estimated from the C-V curve, as illustrated in Figure 3.2(b) and (c), respectively. The extraction of relevant parameters from these distorted C-V curves is therefore troublesome and doubtful. There is therefore a strong need to review the C-V measurement methodology for devices with an ultra-thin gate dielectric.

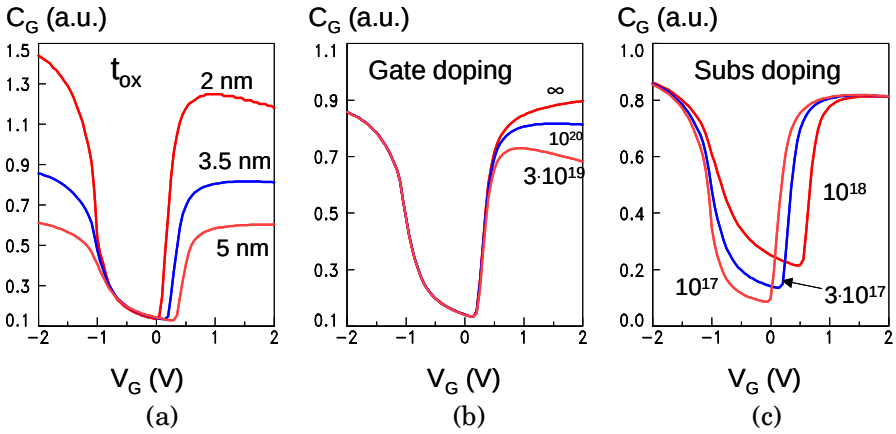


Figure 3.2: *C-V characteristics modeled by R. van Langevelde using MOS Model 11 and an 0.18 μm CMOS technology: the gate dielectric thickness (T_{ox}) as well as the polysilicon gate and substrate doping can be easily extracted from C-V curves.*

3.1.2 Chapter overview

In this chapter, the practical problems in C-V characterization arising when a MOS capacitor shows significant gate leakage current through the gate dielectric are discussed. This chapter starts with a description of the high-frequency (HF) C-V measurements and a discussion on the impact of gate leakage current on C-V measurements (section 3.2). Recommendations on the HF C-V measurement

procedure are then given. Achieved C-V measurements under high gate leakage current on various test structures are also presented. In chapter 3.3, the radio frequency (RF) C-V measurements procedure is described and some RF C-V characteristics are presented.
Details on the design of a test structure dedicated to the measurement of C-V characteristic under high gate leakage current are discussed in chapter 3.4. A new layout is proposed for both HF and RF C-V measurements.
In chapter 3.5, a comparison of some models used to extract parameters such as the EOT is presented. Based on our present knowledge of C-V measurements, we summarize in chapter 3.6 the best measurement procedure and test structure design to obtain reliable C-V curves of MOS structures with ultra-thin oxynitride gate dielectrics. A benchmark of the HF and RF C-V characteristics is also presented. Recommendations for future work are finally given.

This chapter is not a general introduction to C-V measurements. We assume that the reader is familiar with the basic concepts of C-V measurements. The theory of C-V curves can be looked up in the literature: see for example references [4]-[8]. In this chapter, we focus on the additional problems caused by leakage current.

3.2 High-frequency C-V measurements

3.2.1 Measurement description

High-frequency (HF) C-V measurements are carried out using a probe station and an LCR meter: HP 4284A in our case. A DC bias voltage is set across the capacitor, and a small high frequency AC signal is added to it. Typically, the high frequency signal has an amplitude of 10-50 mV and a frequency range of 10 kHz up to 1 MHz. An LCR meter supplies the above mentioned bias, and measures the AC component of the current as a function of time. The amplitude of the AC current gives the impedance (Z) of the device under test; the phase shift (φ) between the voltage and the current oscillations indicates whether the tested device is capacitive (φ=90^o), resistive (φ=0^o), or inductive (φ=-90^o) in nature, as shown in Figure 3.3(a).
The accumulation part of the C-V curve is very similar to that of the quasi-static measurement; be it that slow interface states, that cannot follow the high frequency signal, are not visible. The inversion part is only observed when supplies of minority carriers are present in the form of diffusion edges (source/drain edges) around the capacitor. The substrate cannot generate and recombine minority carriers fast enough to supply the inversion charge. In Figure 3.3(b), a high frequency C-V curve is shown for a NMOS gated diode (capacitor with an active area to provide minority carriers) with a 2 nm oxynitride gate dielectric.

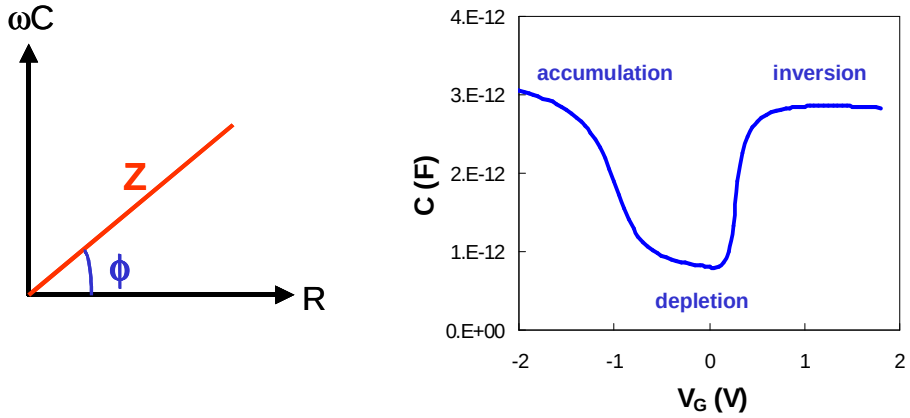


Figure 3.3:*(a) Schematic of the measured impedance (Z) and phase (ϕ). (b) 1MHz high frequency C-V curve for an NMOS gated diode with the accumulation, depletion and inversion regions highlighted.*

3.2.2 Impact of high leakage current on C-V measurements

3.2.2.1 Problem description

The impact of high gate leakage current on C-V measurement is illustrated in Figure 3.4(a). C-V characteristics measured on NMOS gated diodes with thin gate dielectrics having a gate leakage current density of 1 and 100 A/cm^2 (at V_{GS}=1 V) are compared. The capacitance for the gated diode with the leakiest (also the thinnest) gate dielectric begins to sharply decrease in the accumulation and inversion regions, that is to say under strong gate bias. The mechanism responsible for the capacitance attenuation is gate tunneling current as well as series resistance (R_S) such as the gate, channel and bulk resistances [9-12]. The value of R_S strongly depends on the layout of the test structure and sheet resistivity of the substrate.

The tunneling conductance (G), intrinsic capacitance (C) and R_S are defined in Figure 3.4(c) and represent an equivalent circuit model for the MOS device under high leakage current. However, among the equivalent circuit models available on the LCR meter, none of them takes into account both G and R_S. HF C-V measurements are performed using the parallel circuit model (Figure 3.4(b)) which takes into account G but neglects R_S. Using impedance transformations, the measured capacitance (C_m) and conductance (G_m) can be expressed as a function of G, C and R_S (Equation 3.3).

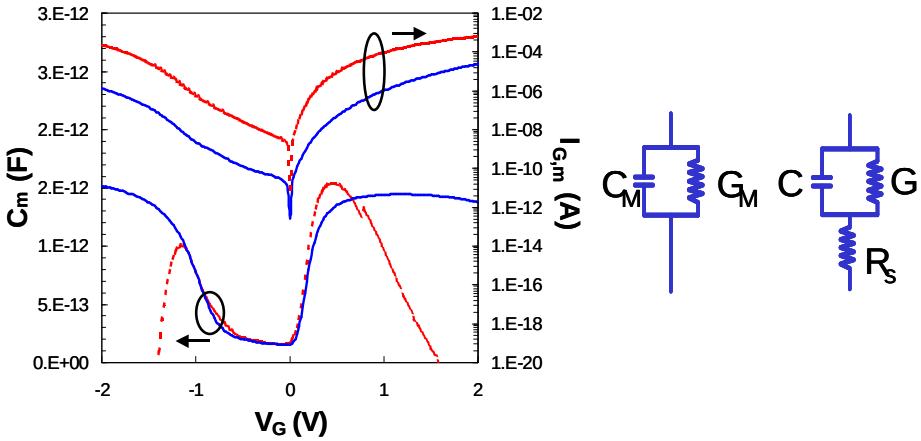


Figure 3.4:*(a) Comparison of C-V characteristics measured on NMOS gated diodes with thin gate dielectrics having a gate leakage current density of 1 and 100 A/cm^2 (at V_{GS}=1 V). Small signal equivalent circuit model of MOS capacitor: (b) parallel measurement model and (c) model including G and R_S.*

$$C_m = \frac{C}{(GR_s+1)^2 + \omega^2 C^2 R_s^2}$$
$$G_m = G.\frac{(1+R_sG)+(C\omega)^2 R_s/G}{(1+R_sG)^2+(R_sC\omega)^2} \qquad (3.3)$$

This relationship (3.3) indicates that G_m and C_m are affected by R_S. Moreover, G and R_S modify the measured capacitance, whatever the frequency used. If the leakage current is high (i.e. G is high), the measured capacitance is decreasing and the effect of R_S becomes amplified. It should be also noted that high gate leakage associated with high R_S will have an impact on the actual voltage across the oxide: it will be less than the externally applied voltage. From our experimental data, this difference is marginal and therefore was not taken into account in this report.
Furthermore, the LCR will not be capable to accurately measure the capacitance if the tunneling leakage current is too large. Indeed, the dissipation factor (D), as defined by the HP 4284A operating manual [13] when using the parallel equivalent model (Figure 3.4(b)), indicates that the measurement error will strongly increase when D and therefore G increase (3.4), as illustrated in figure 2.2.1.2.

$$D = \frac{G_m}{2\pi f C_m} \quad \text{and} \quad \%error = 0.1\sqrt{\left(1+D^2\right)} \qquad (3.4)$$

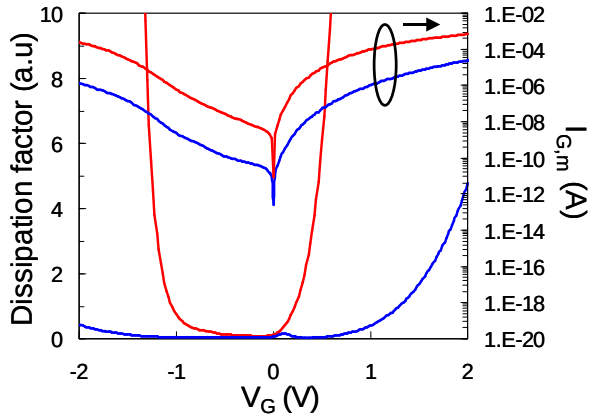


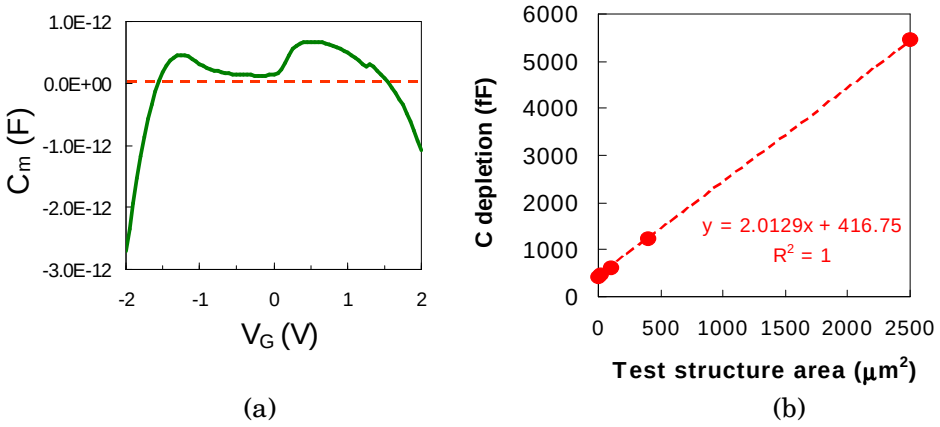
Figure 3.5: *Dissipation factor measured on NMOS gated diodes similar to Figure 3.4(a). High gate leakage current increases strongly the dissipation factor and therefore the measurement error.*

As a result, high frequencies should be used to measure MOS devices having ultra-thin leaky gate dielectrics in order to minimize the instrumentation error. However, an important drawback of higher frequency measurements is the larger sensitivity to

signal transfer loss and residual inductance. This problem can be solved with a radio-frequency measurement methodology (see chapter 3.3).

3.2.2.2 Correction for parasitic elements

C-V measurement on devices with ultra-thin gate dielectric are very sensitive to parasitic elements. Indeed, in order to reduce the DC component (namely the gate leakage current), small area test structures are used yielding small measured capacitances. As a results, parasitic elements, arising from the measurement set-up (e.g. cables) or/and from the test structure (e.g. connections), are no more negligible. The measurement of negative capacitance is the consequence of these parasitic elements, as shown in figure 2.2.2.1(a).



(a) (b)

Figure 3.6 : *(a) C-V measurements under high gate leakage current. Negative measured capacitances are the consequences of parasitic elements that are not taken into account in the equivalent MOS scheme (**Figure 3.4(b)**). (b) Extracted pad capacitance from capacitance in depletion versus the gate area. The depletion capacitance corresponds to the minimum capacitance measured on the square gated. The pad capacitance extracted is C_{pad}~417 fF.*

A parasitic element that will limit the downscaling of the test structure area is the gate-substrate capacitance, also called pad capacitance (C_{pad}), which is in parallel with the MOS structure (see Figure 3.7). As the gate capacitance becomes smaller, errors in estimating the parasitic pad capacitance can results in significant errors in calculating or estimating the oxide thickness.

The pad capacitance is extracted by measuring the gate capacitance in depletion (not affected by *G*) on square gated diodes with different gate areas but same pad interconnect scheme. The measured capacitance is plotted versus gate area (*A*) and the pad capacitance is found from the intercept with *A*=0 axis. The capacitance in depletion of square gated diodes was measured and a pad capacitance of 417 fF was extracted, as shown in Figure 3.6(b).

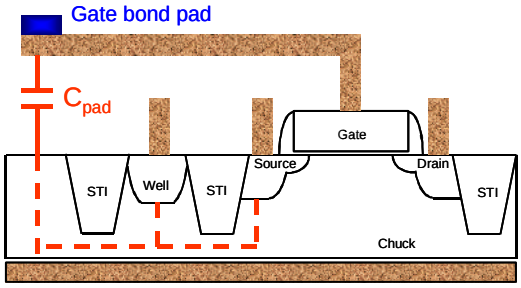


Figure 3.7: *Schematic local MOS with a parasitic pad capcitance.*

In order to verify this value, the pad capacitance has been estimated from a simple calculation of the field oxide and pre-metal dielectric (PMD) capacitances. For the technology used to fabricate 40 nm transistors, the field oxide layer and the PMD layer have a thickness of respectively 365 and 765 nm. The dimensions of the bond pads are 100×100 μm^2. The resulted capacitances of the field oxide and PMD layers are respectively 450 and 960 fF, which results in a total capacitance (adding the two capacitances in series) of about 306 fF. The difference between the measured and estimated pad capacitances can be attributed to the variations that can occur in the thickness of the field oxide and PMD layer during the process. Yet, the two values are in the sub-pico Farad regime which will limit the downscaling of the dimensions of the test structure needed for devices with high leakage gate dielectric.
It is therefore essential to minimize this parasitic pad capacitance, reducing its area, working on isolated MOS structures having eventually a triple well implant and using an optimized MOS design to avoid uncertain corrections (see section 3.2.5).

3.2.2.3 Corrections for external resistances (G, R_s)

If the test structure is not correctly designed, an equivalent scheme taking into account the tunneling current as well as the series resistance should be applied to correctly interpret the C-V measurements. Among various correction models proposed in the literature, two have been investigated carefully.

- *Dual frequency method [14]*

A method using C-V measurements at two frequencies in parallel mode has been proposed [14]. This method is simply equating the parallel and accurate models (Figure 3.4(b) and (c)); the use of two frequencies allows to extract the intrinsic oxide capacitance (C), G and R_S. This method does not impose a theoretical limit on the choice of the two frequencies. However, the measurement error increases when the gate leakage current is increasing (see equation 3.4) and the frequency range where this method can be applied is therefore shifting towards higher frequencies reaching a limit when considering the frequencies available on the LCR meter (maximum frequency is 1 MHz).
A good illustration of this problem is shown in Figure 3.8 where the dual frequency method has been applied on an NMOS device with an ultra-thin oxynitride (1.5 nm EOT, J_G=10 A/cm^2 at V_{GS}=1 V) using three different frequency ranges: 1 MHz - 400

kHz, 1 MHz – 100 kHz, 400 kHz – 100 kHz. The measured C-V characteristics are plotted in Figure 3.8(a), (b) and (c) with underlined parts of the curves where *D* is below 1 or 0.1. As defined in (3.4), *D* is a good figure of merit for the measurement error. The measurement error is increasing with decreasing the frequency yielding to capacitance values measured with a high measurement error. The resulted corrected C-V curves are very different depending on the chosen frequency range and, in our case, the obtained "corrected" C-V curves do not result in the expected intrinsic C-V characteristic. We will show later that this method can be successfully applied on C-V characteristics measured at RF frequencies where the measurement error is small (chapter 3.3).

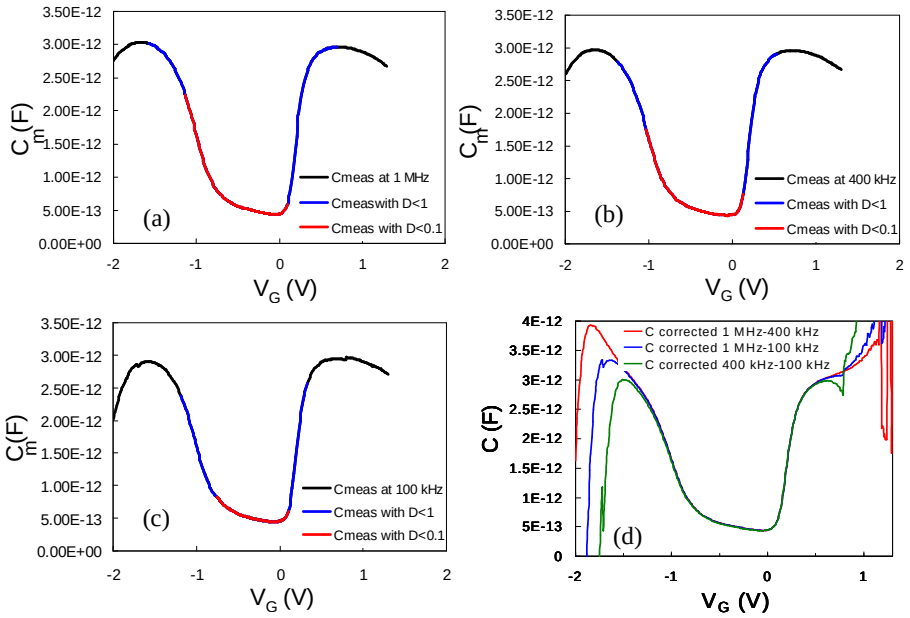


Figure 3.8: *Measured C-Vs on NMOST with a 15 Å EOT gate dielectric (J_G=10 A/cm^2 at V_{GS}=1 V) at various frequencies: (a)1 MHz, (b)400 and (c)100 kHz. The capacitance measurements for various D are underlined. (d) Corrected C-V curves using the dual frequency method [11] for various frequency. The corrected C-V curve is dependent on the frequency ranges. The obtained "corrected" C-V curves do not result in the intrinsic capacitance.*

– *Transmission line based method (Barlage's method) [15]*

Another method, valid only in inversion, has been proposed by Barlage et al. [15]. Note that due to the lack of accurate fitting model to extract EOT in the inversion part of the C-V curve, only the CET_{inv} will be measured on the C-V obtained with this method. This approach is based on the application of a lossy transmission line model to the MOS transistor. The channel is partitioned in segments, each segment being a 3-component equivalent circuit as defined in Figure 3.4(c). A correction factor is calculated based on the sheet resistance of the inversion channel (called r_s)

and on the normalized tunneling impedance of the oxide (called r_t). An advantage of this method is that the correction factor is independent on the measured capacitance. It depends only on DC measurements that are more accurate than C-V ones. However, since the MOS structure is characterized by the parameters r_t and r_s along with the channel length, the model must be restricted to cases where the lateral potential drop along the channel is very small compared to the gate-to-channel voltage. If the gate leakage current, flowing along the lateral channel to source/drain, produces a significant voltage drop, then r_t and r_s become position dependent as well as bias dependent and the problem becomes much more difficult. This places a restriction on the gate length and the gate leakage current for which the model can provide accurate correction. The channel potential has been predicted for long channel device with high gate leakage current using MOS Model 11 (MM11) [16]. The influence of the gate leakage current on the channel potential is not negligible for long channel device with high leakage current, as presented in Figure 3.9.

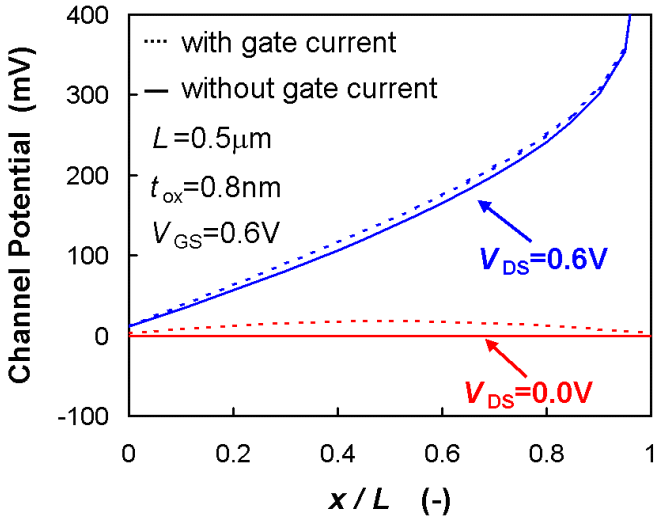


Figure 3.9: *Influence of the gate current on the channel potential distribution for a device with a channel length of 0.5 μm and an oxide thickness of 0.8 nm, as predicted by MM11 [16]. For V_{GS}=0.6 V, the gate current was 85 μA.*

This correction method [15] has been applied on 10, 3 and 1 μm channel length NMOS transistors having various thin gate oxynitride dielectrics. The calculated characteristics of r_t and r_s exhibit similar behavior independently of the gate dielectric (i.e. independently of the gate leakage current). In Figure 3.10, examples of the obtained r_t and r_s characteristics are shown for various gate lengths.

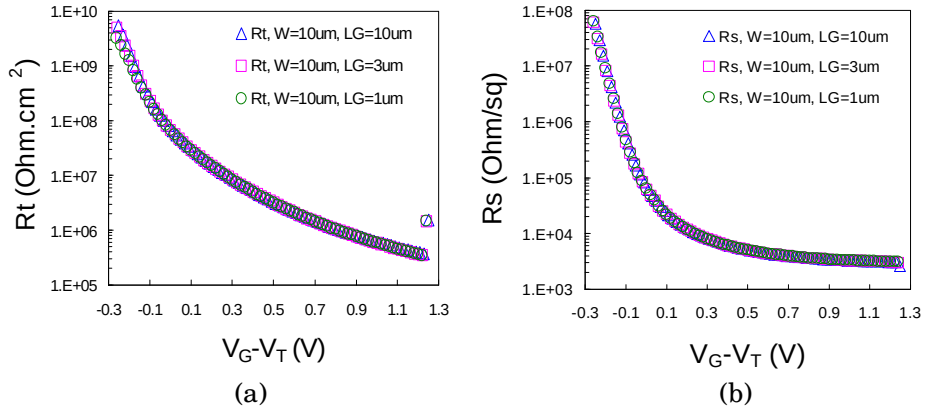


Figure 3.10: *Calculated (a) tunneling resistance (r_t) and (b) channel resistance (r_s) according to [15] on NMOS transistors with a 10, 3 and 1 µm gate length.*

The calculated r_t is decreasing exponentially when increasing gate bias independently of the channel length. This behavior can be explained by the fact that the gate leakage current is exponentially increasing with gate bias due to direct tunneling mechanism. Moreover, the calculated r_s characteristics show, as expected, an exponential decrease at small gate biases (linear regime) followed by a linear decrease when the saturation regime is reached. R_S is also independent to the channel gate length.
Although the r_t and r_s calculated characteristics exhibit normal behavior when varying the gate bias, the resulting corrected capacitance is not the expected intrinsic capacitance. For devices with a rather "thick" gate dielectric (EOT = 20 Å, J_G=0.01 A/cm^2 at V_{GS}=1 V), the measured and corrected capacitances are overlapping, as presented in figure 2.2.3.4(a). Indeed, the measured capacitances are already leakage-free and therefore do not need to be corrected. Nevertheless, it is interesting to underline that the calculated capacitances are independent of the channel length. For devices with ultra-thin gate dielectrics (EOT ≤ 14 Å, J_G=10 A/cm^2 at V_{GS}=1 V), the calculated capacitances are not free of all external resistance as they are no more independent of the channel length (figure 2.2.3.4(b)). The obtained "corrected" C-V curves are still attenuated in inversion showing that they are not leakage free.

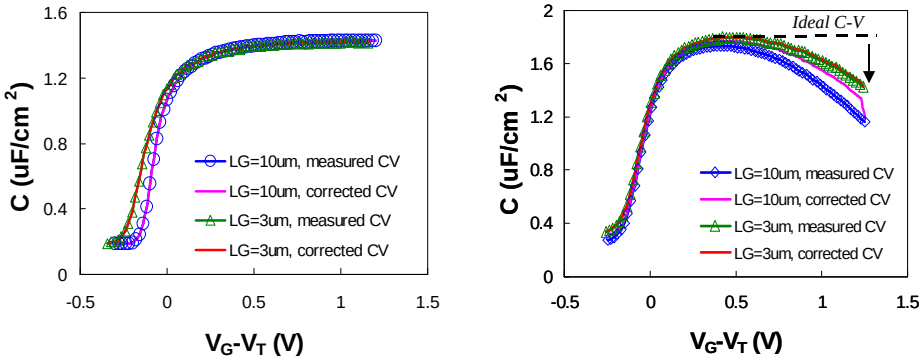


Figure 3.11: *Measured CV characteristics for NMOS devices with a (a) 20 Å (J_G=0.01 A/cm² at V_{GS}=1 V) and (b) 14 Å (J_G=10 A/cm² at V_{GS}=1 V) oxynitride gate dielectric. Corrected CV curves using the Barlage's method [12] are also plotted. (b) The obtained "corrected" C-V curves are not leakage/resistance free.*

- *MOS model 11 [16]*

As discussed earlier in this section, the assumption of the Barlage's method that r_t and r_s are uniform along the channel is not valid when the gate leakage current produces a significant lateral voltage drop which is the case of the studied ultra-thin oxynitride films. Moreover, other external resistances might need to be taken into account on top of the channel and tunneling resistances. To verify this hypothesis, MOS model 11 (MM11) was used to model the capacitance in inversion of a 10x10 µm² NMOS with highly leaky dielectric (J_G=100 A/cm² at V_{GS}=1 V).

MM11 is an advanced compact model developed by Philips [17]. It is a symmetrical, surface-potential-based model. It includes an accurate description for all physical effects important for modern and future CMOS technologies such as gate tunneling current [15].

First, the NMOS channel was partitioned in 10 segments (each segment being a 3-component equivalent circuit similar to the one described in Figure 3.4(c)). For each of these segments, the tunneling and series resistances are bias dependent. This model is more accurate than the Barlage's correction method as it includes the channel non uniformity under high gate leakage current (Figure 3.9). This segmentation yields a decrease of the capacitance in inversion but does not explain the total decrease of the measured capacitance, as shown in Figure 3.12(a). When adding an extra resistance in series with the gate terminal, the modeled C-V curve is describing the roll-off of the measured C-V observed around the threshold voltage, as illustrated in Figure 3.12(b). This extra resistance could be the gate resistance resulting from a too wide and/or to the lack of close and sufficient contacts. In order to obtain a good fit with the measured C-V curve, the value of this resistance was set at 400 Ω which is a rather high value. Yet, for large positive gate bias (strong inversion), the modeled C-V is not reproducing the large decrease of the measured C-V towards negative values. In order to describe the negative values of the measured

capacitance in strong inversion, an inductance was added in series that could be the result of remaining parasitic inductance (cabling, pad capacitance etc…) that have not been properly corrected. The value of this inductance was set to 400 nH, which is a very high value and shows that de-embedding structures are strongly required for these meaurements. As shown in Figure 3.12(c), the modeled and measured C-V characteristics are well correlated in inversion.

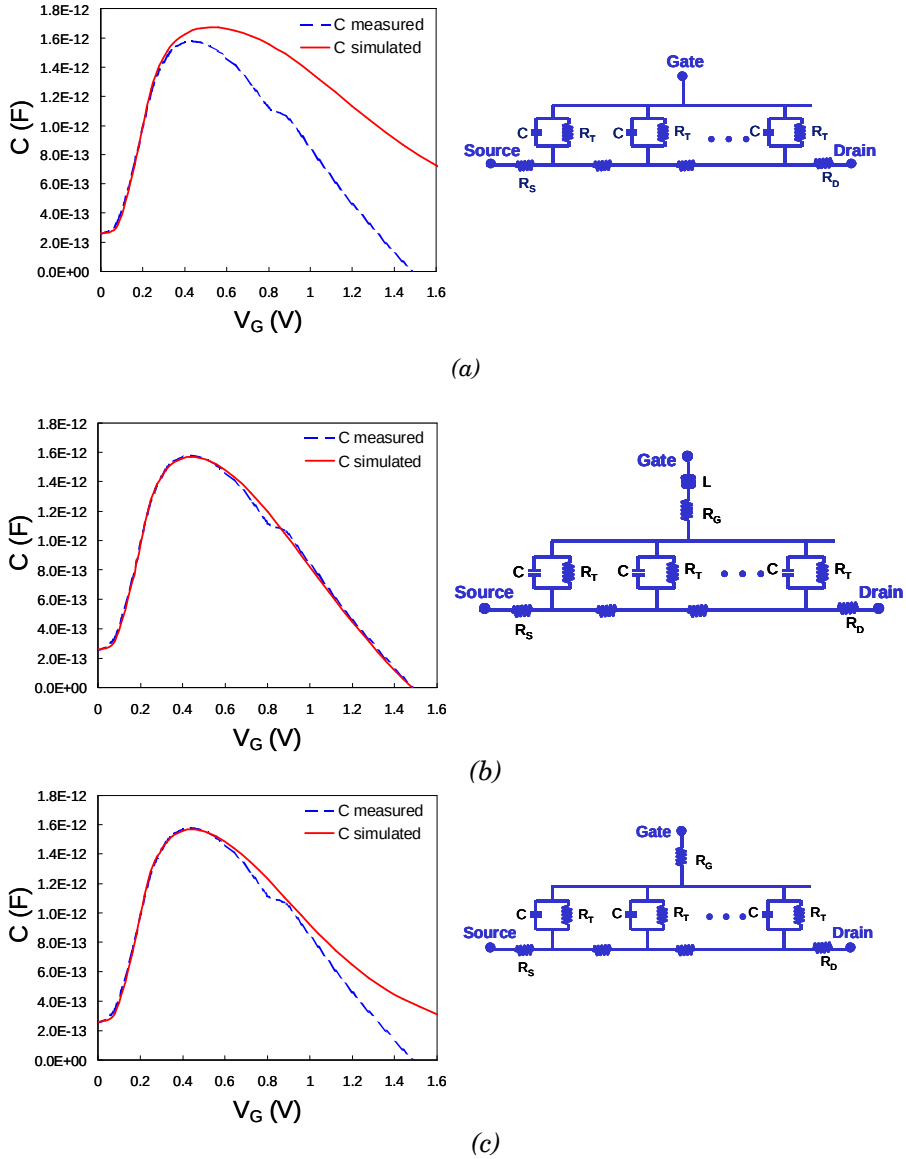


Figure 3.12: *Measured and modeled (using MM11 where tunneling and series*

resistances are bias dependent) of C-V characteristics in inversion for a 10x10 μm^2 NMOS transistor with an ultra-thin gate dielectric (J_G=100 A/cm^2 at V_{GS}=1 V). A transmission line model for the MOSFET channel has been implemented: 10 segments have been introduced (a). In order to explain the large attenuation of the capacitance for large positive gate bias (inversion regime), a gate resistance (b) and an inductance (c) need to be added in series.

To summarize, MM11 predicts accurately the C-V measurement degradation behavior observed under high gate leakage current. The non uniformity of the channel potential under high gate leakage current plays an important role in the modellisation of C-V curves.
Because it can model successfully C-V characteristic of leaky oxide, MM11 was finally used to simulate the full C-V curve and to correct for parasitic elements and external resistances. The resulted simulated C-V curve reproduces nicely the intrinsic C-V characteristic of a device with an ultra-thin leaky gate dielectric (J_G=100 A/cm^2 at V_{GS}=1 V), as presented in Figure 3.13.

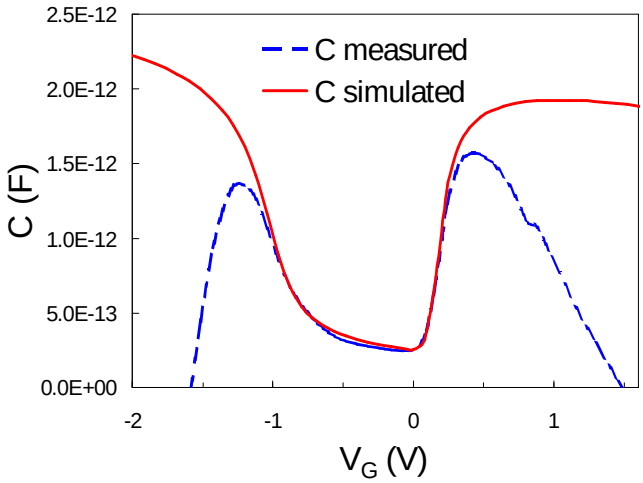


Figure 3.13: *Measured and modeled (using MM11) of C-V characteristics for a 10x10 μm^2 NMOS transistor with an ultra-thin gate dielectric (J_G=100 A/cm^2 at V_{GS}=1 V).*

3.2.3 High frequency C-V measurement procedure

A typical cabling scheme is presented in Figure 3.14 for a transistor as the test structure.
Since the measurement of devices with an ultra-thin oxide requires very small signals (reduced test structure area to decrease the DC component), parasitic effects become

non negligible. A full calibration of the measurement set-up is of first importance in the measurement procedure.

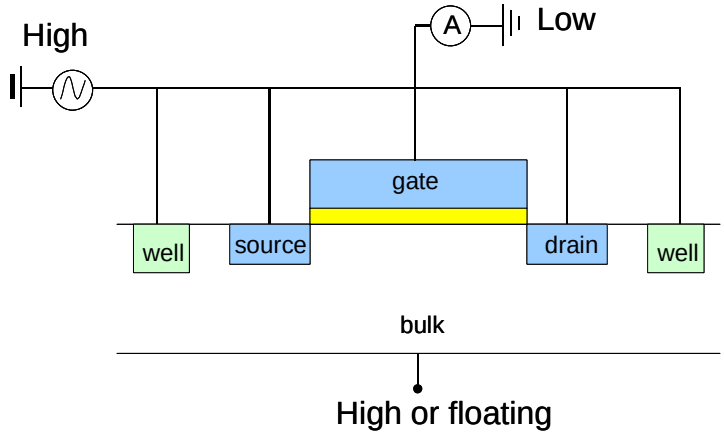


Figure 3.14: *Schematic of transistor test structure used for C-V measurements. The chuck can be connected to the High terminal or being floating.*

The basic procedure steps that should be executed prior to the actual C-V measurements and whatever the test structure, are the following:

1) Cleaning of the probes

It is of first importance to always start a measurement with clean probes. This will ensure good contacts and will result in reliable and reproducible measurements.

2) Measurement of the contact resistance

Some variations in the processing might have a direct impact on the contact resistance (e.g. silicidation process). This resistance should be measured (e.g. on Kelvin test structure) prior to C-V measurements and should be below 10 $\Omega.cm^{-2}$.

3) Calibration of the measurement set-up

In order to calibrate the LCR meter and correct for parasitic elements such as series impedance induced by the cables and parallel parasitic capacitances, OPEN and SHORT corrections are required. The cables length should be kept as short as possible and the connections should be fixed during the measurement. Also, ideally, the needles should be handled automatically to avoid any variation of the parasitic impedance of the set-up.

Some further corrections for parasitics and external resistances might be needed depending on the gate leakage current level and on the test structure. The presence of dedicated OPEN/SHORT test structures or clever design tips on the test will prevent extra correction for parasitic capacitance and external resistance. However, most of the test structures present on current available masksets require a measurement of the gate pad capacitance that should be subtracted to the measured capacitance.

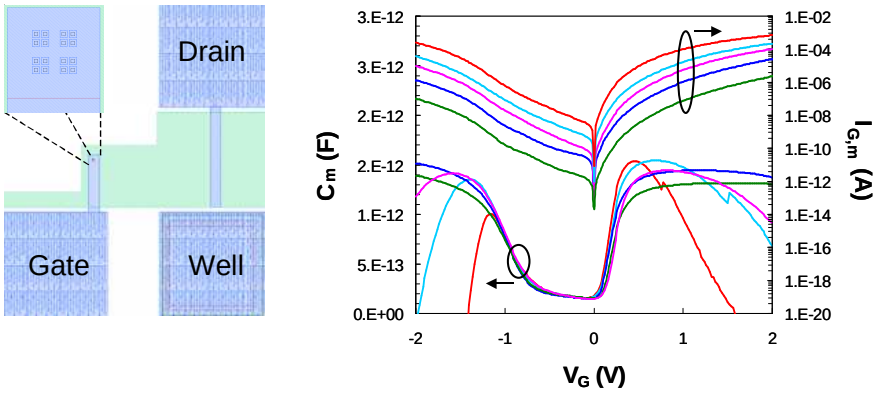
3.2.4 HF C-V measurements: 1 MHz

C-V characteristics of MOS devices with ultra-thin gate dielectrics have been measured at 1 MHz (maximum frequency possible on the HP4284A) on various test structures. The main results for each studied test structure have been reported below.

– *Square gated diodes*

Figure 3.15(a) shows the layout of a square gated diode. A large number of these gated diodes have been drawn with various areas having a common well and active areas.

As mentioned in section 3.2.2.2, the large parasitic pad capacitance limits the down scaling of the area of the test structures. It was found that for a 5x5 μm^2 gated diode, the pad capacitance was already more than 40 % of the total measured capacitance. Therefore the minimum area for these gated diode structures yielding accurate C-V measurements is 10x10 μm^2. For this specific test structure, accurate C-V measurements can be obtained for NMOS devices with a gate leakage current density below 10 A/cm^2 (at V_{GS}= 1V), as presented in Figure 3.15(b).



(a) (b)

Figure 3.15: *(a) layout of the square gated diode sructure. (b) Measured C-V and corresponding gate leakage current (I_G) characteristics for five NMOS devices with thin oxynitride gate dielectrics. The thickness of the oxynitride films was varied from 2 down to 1.2 nm EOT (represented with different colors). The measurements were performed on the 10x10 μm^2 gated diode test structure available on the MINOXG maskset. Accurate C-V curves are obtained for devices with a gate leakage current density below 10 A/cm^2 at V_{GS}=1 V.*

For this test structure, the well contact was drawn too far from the device itself, inducing a high resistance in accumulation. Therefore, the substrate contact was taken at the backside of the wafer. In order to reduce the bulk resistance, epi wafers were used. Epi wafers yield lower R_S (~ factor 4 decrease as compared to bulk

substrate) resulting in a better C-V characteristic in accumulation (capacitance attenuation starts at higher gate bias), as presented in Figure 3.16. Note that the doping concentration of the epi bulk Si was around 5×10^{17} at/cm^3 in this experiment. This bulk doping concentration could be increased for a further reduction of the bulk resistance. Moreover, the bulk resistance can also be reduced by designing a large amount of bulk contacts close to the device itself (see section 2.5).

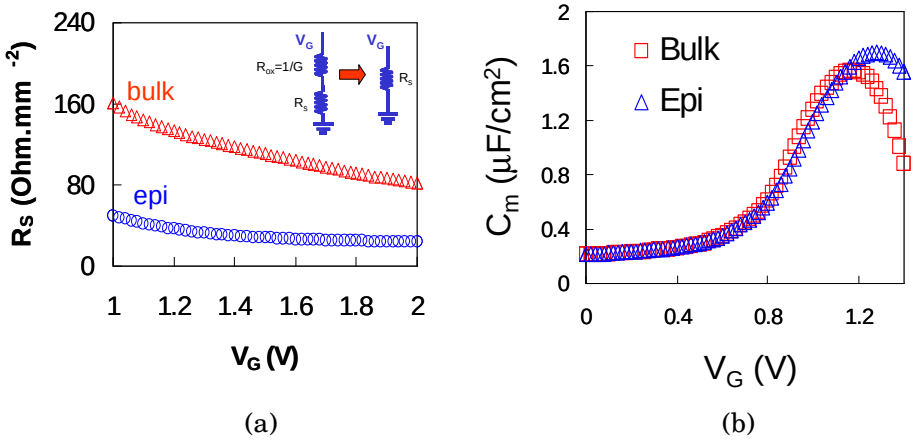


Figure 3.16: *(a) R_S and (b) C-V curves measured in accumulation on PMOS gated diode having an ultra-thin gate dielectric and formed on an epi or standard bulk wafers. R_S was extracted from DC measurements on a large (0.81 mm^2) capacitor without drain edge in order to eliminate any parallel leakage path (e.g. gate to drain leakage) which could interfere with the measurement of R_S. Because of the high gate leakage, the oxide resistance (R_{ox}) can be neglected as shown in the equivalent circuit in the inset of the plot. C-V characteristics were obtained from 10x10 µm^2 gated diode.*

– *2 transistors in parallel (matching test structure)*

A structure consisting of two transistors has been studied. Each transistor has a gate width and a gate length of 10 µm, yielding a total area of 20x10 µm^2 (Figure 3.17(a)). As shown in Figure 3.17(b), accurate C-V measurements can be measured on NMOS transistors with a maximum gate leakage current density below 10 A/cm^2 (at V_{GS}=1 V).

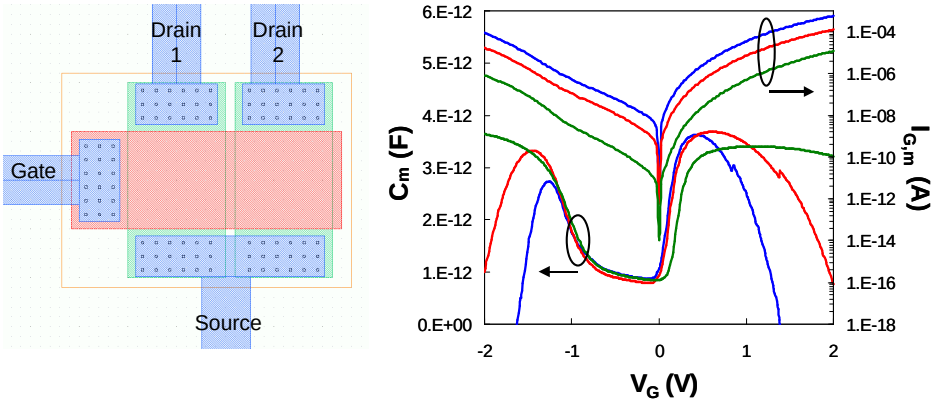


Figure 3.17: *(a) Layout of the 2-transistorsmatching structure. (b) C-V and corresponding gate leakage current (I_G) characteristics for NMOS devices with thin oxynitride gate dielectrics. The thickness of the oxynitride films was varied from 2 down to 1.2 nm EOT (represented with different colors). The C-V measurements were performed on two 10x10 μm^2 transistors in parallel. The gate leakage current measurements have been performed on NMOS transistors having an area of 40 μm^2. Accurate C-V curves are obtained for devices with a gate leakage current density below 10 A/cm^2 at V_{GS}=1 V.*

– *Larrays transistors*

The transistors have a fixed gate width and variable gate length and are designed with common source and well. Drain and gate are individually contacted, as shown in Figure 3.18. C-V measurements have been performed on NMOS transistors having an ultra-thin oxynitride gate dielectric (CET_{inv}=20.9 Å and J_G=5 A/cm^2 at V_{GS}=1 V) and a gate length of 10, 3 or 1 µm, as shown in Figure 3.19(a). The normalized capacitances are independent of the channel length showing that for these transistors the effect of external series resistances is negligible. The discrepancies observed for the C-V curve measured on the 1 µm channel length can be attributed to an inaccuracy of the channel length measured on top SEM images or/and to the noise level that becomes non negligible when measuring such small capacitances. However, having too large gate length will increase the gate leakage current component which will attenuate the capacitance, as shown in Figure 3.19(b) for an NMOS transistor with a thin gate dielectric (J_G=18 A/cm^2 at V_{GS}=1 V). A gate length of 3 µm was chosen to investigate the maximum gate leakage current allowed to get accurate C-V curve using NMOS transistor as a test structure. NMOS transistors with a gate dielectric having a maximum gate leakage current density of 20 A/cm^2 yield accurate C-V measurements, as illustrated in Figure 3.20.

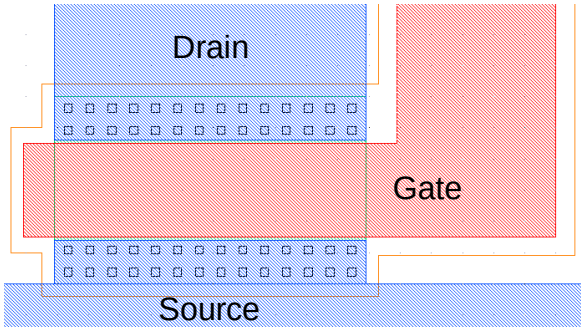


Figure 3.18: *Layout of the Larray transistors.*

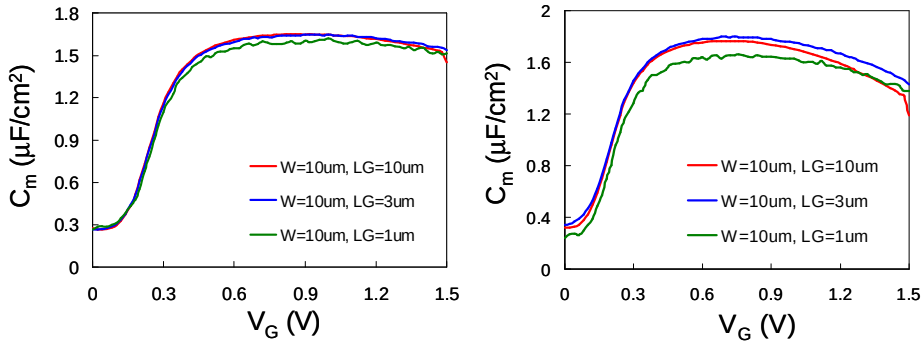


Figure 3.19: *C-V curves measured on NMOS transistors with a gate length of 10, 3 and 1 µm. The gate dielectric has a* CET_{inv} *of (a) 20.9 Å (*J_G*=5 A/cm*^{2} *at* V_{GS}*=1 V) and (b) 19.7 Å at* V_{GS}*=1 V (*J_G*=18 A/cm*^{2} *at* V_{GS}*=1 V).*

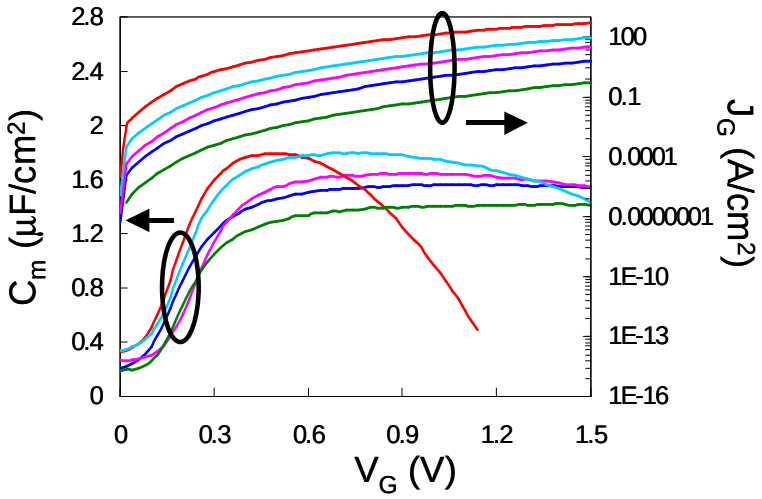


Figure 3.20: *C-V curves measured on NMOS transistors with a gate length of 3 µm and related gate leakage current density (J_G). . The thickness of the oxynitride films was varied from 2 down to 1.2 nm EOT (represented with different colors). Accurate C-V curves are obtained for devices with a gate leakage current density below 20 A/cm².*

3.3 Radio-frequency C-V measurements

3.3.1 Measurement description

As mentioned in chapter 2, the measured capacitance is affected by gate leakage current (G) and by the external resistance (R_S). Increasing the frequency of the measurement will not help if R_S is high (equation 3.3). However, increasing the measurement frequency is required to minimize the instrumentation error (equation 4).

The radio-frequency (RF) measurements are carried out on optimized test structures (see section 3.3.3) using a Cascade probe station and a network analyzer (HP 8510C). From S parameter measurements, Y parameters are determined and open/short de-embedding is carried out [18-20]. Figure 3.21 shows an example of the Y_{11} parameter (gate admittance) of a NMOS device with high leakage current (J_G=50 A/cm^2 at V_{GS}=1 V). The imaginary component (ωC) dominates only for a certain frequency range, in this particular case between 20 MHz and 2 GHz. Note that this effect will not be observed if the correct measurement model was applied (here the measurement circuit model C parallel to G was used, as described in Figure 3.4(b)). The real component part attributed to the gate conductance (i.e. leakage current component) will drastically change with the gate bias.

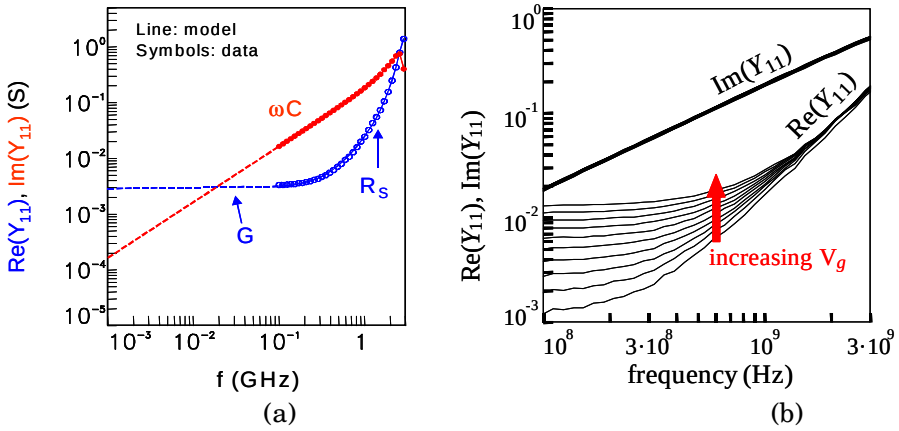


Figure 3.21: *Measured real and imaginary components of the input admittance of a NMOS capacitor in inversion. The device has 1872 μm^2 gate area and an ultra-thin gate dielectric (J_G=50 A/cm^2 at V_{GS}=1 V). (a) the imaginary component (ωC) dominates only for frequencies between 20 MHz and 2 GHz. (b) The gate bias was varied between 0.3 V and 1.2 V in steps of 100 mV. The imaginary parts almost overlap, but due to the exponential increase of gate leakage current with V_G, the real part of Y_{11} changes drastically.*

The minimum frequency is dictated by the magnitude of the gate leakage. The highest measurement frequency is constrained by the external resistance (R_S). When excessive gate leakage forces the application of RF measurement frequencies, R_S must therefore be reduced to a minimum to still allow a good measurement of capacitance. The selection of the correct measurement frequencies is therefore crucial and strongly depends on the test structure used. The measured frequency range should be chosen in the region where C is frequency independent (i.e. where the imaginary component dominates the real one) [21]. It is therefore important, prior to the actual C-V measurements, to measure the capacitance as a function of the frequency and at several gate bias since G and R_S will have a different behavior at different gate bias. In Figure 3.22(a), the measured capacitance and conductance as a function of the frequency is presented for an NMOS device with an ultra-thin gate dielectric (J_G=100 A/cm^2 at V_{GS}=1 V).

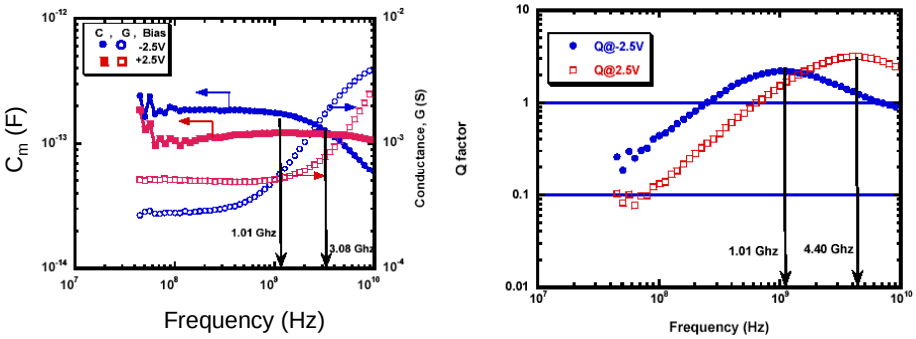


Figure 3.22: *(a) Capacitance (=$|ImY_{11}|/\omega$) and conductance ($|ReY_{11}|$) measured on NMOS transistors with a gate length of 0.5 μm and an ultra-thin gate dielectric (J_G=100 A/cm^2 at V_G=1 V) as function of the frequency. The measurements were done in strong accumulation (V_G=-2.5 V) and strong inversion (V_{GS}=2.5 V). (b) The quality factor (Q) of the same device was calculated for the two gate biases: V_G=+2.5 and –2.5 V. The peaks of the Q factor were reported in (a) and do not correspond to the region where $|ImY_{11}|/\omega$ is flat.*

As expected, the frequency range where the capacitance is independent of the frequency depends on the gate bias. However, there is a common frequency range where the capacitance is independent of the frequency in both strong accumulation (V_{GS}=-2.5 V) and strong inversion (V_{GS}=2.5 V). Choosing the maximum of the quality factor (Q=1/D) does not correspond to this frequency range, as illustrated in Figure 3.22(b). Care should be taken when measuring devices with different gate lengths since the frequency range where C will be constant over the frequency will shift (smaller gate length, higher frequencies), as illustrated in Figure 3.23.

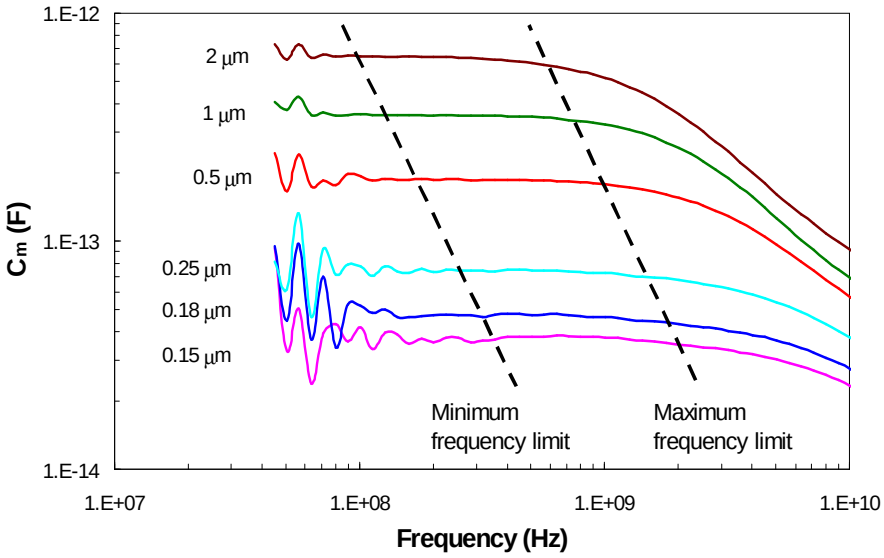


Figure 3.23: *Capacitance measured in strong accumulation (V_{GS}=-2.5 V) as a function of frequencies for NMOS transistors with an ultra-thin gate dielectric (J_G=100 A/cm2 at V_{GS}=1 V). The dashed lines indicate the lower and upper limits of the frequency range where the capacitance is independent of the frequency.*

Once the correct measurement frequencies have been selected, the actual RF C-V measurements can be performed. First, the capacitance is measured as a function of the frequency (chosen using the method described above). Because the measurement model used (Figure 3.4(b)) is not accurate for measurements under high gate leakage current (section 3.2.2), a correction for R_S and G should be applied to get the intrinsic capacitance. Since the instrumentation error is very small at RF frequencies, the dual frequency method (see section 3.2.3) can be applied.

3.3.2 RF C-V measurements

RF test structures (so called HFLarrays) have been measured and consist of two gate fingers of variable gate width and length (Figure 3.24). As presented in Figure 3.25(a), after correction, the resulted C-V curves are frequency independent, showing the accuracy of the method. A small variation is obtained when using smaller gate length (0.5 mm), as shown in Figure 3.25(b). This difference can be attributed to an uncertainty in estimating the physical gate length and to a change in the substrate doping: the pockets starting to influence the channel doping.

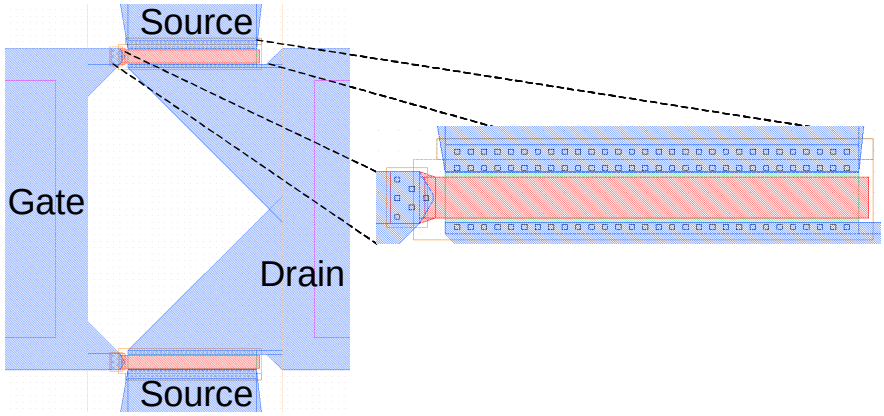
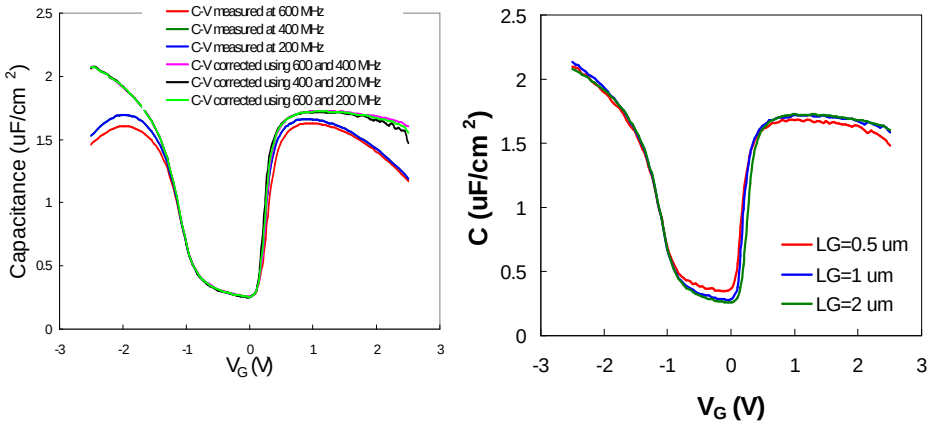


Figure 3.24: *Layout of the HFLarray test structure*



(a) (b)

Figure 3.25: *(a) Measured and corrected capacitances for NMOS transistor with a gate width and length of respectively 20 and 2 μm and an ultra-thin gate dielectric (J_G=100 A/cm^2 at V_{GS}=1 V). The resulted corrected C-V curves are frequency independent, proving the validity of the method. (b) Corrected C-V curves for NMOS transistors with a gate width of 20 μm and variable gate lengths: 2, 1 and 0.5 μm.*

Subsequently, RF C-V measurements on RF structures with multifinger capacitors (in order to reduce the gate resistance) with a gate width of 2.6 μm and a gate length of 0.2 or 1 μm (Figure 3.26). The total gate area on active area is 1872 μm^2 for all devices. Dedicated OPEN and SHORT test structures are also available. However, as a result of an error in the SHORT structure (the shortening was not done on the device level but on the contact one), an extra inductance has to be added in series to

the equivalent circuit model presented in Figure 3.4(c). This inductance could be fitted and was found to be about 80 pH. Furthermore, since these devices had a high well resistance the capacitance in accumulation and depletion could not be accurately determined. Therefore, only the capacitance in inversion was studied. With the four-element model, the external resistance could also be fitted, and after correcting the *Y* parameters for this external resistance and the parasitic inductance, the capacitance was calculated using $C=\mathrm{Im}(Y_{11})/\omega$. Subsequently, we obtained the intrinsic capacitance from the capacitance difference between a 1 m and a 0.15 m gate ength device, and this is shown in Figure 3.27.

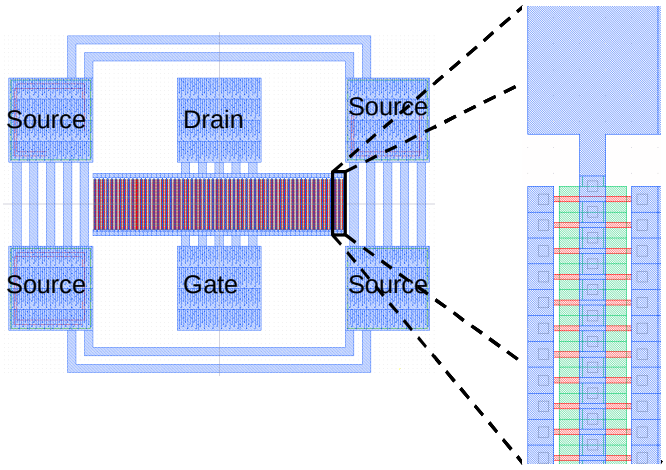


Figure 3.26: *Layout of the MINOXG RF test structure*

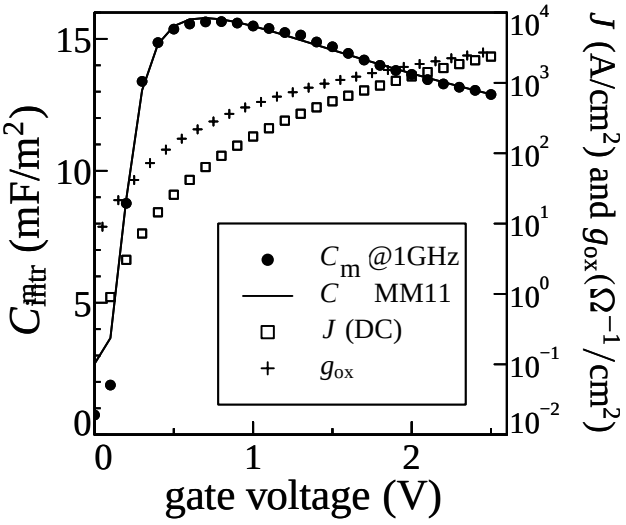


Figure 3.27: *Inversion capacitance (C) and gate-leakage current density of a high-leakage dielectric. The inversion capacitance is correctly measured even when the gate-current density exceeds 1 kA/cm^2, as confirmed by the excellent fit with MM11.*

3.4 Design considerations

As mentioned in previous sections, G and R_S have a strong impact on the measured capacitance. Some optimizations in the design of the test structure can be done to reduce the impact of these two elements. Also, the parasitic elements should be accurately known to perform accurate de-embedding.

3.4.1 Reduce external resistance

Rs has to be minimized in inversion and accumulation, calling for specific design considerations:

- In inversion

The inversion layer is the dominant external resistance [17] in the inversion regime. A wide, short capacitor has advantageous dimensions to drain the inversion charge from the inversion layer. To keep the gate resistance low, not too wide gates must be designed and several contacts should be drawn close to the device.

- In accumulation

In accumulation, the bulk resistance is dominating. A straightforward way to reduce the bulk resistance is to short the active and well areas: no STI between the well and source/drain regions. This should be done in a symmetrical way (well/source and well/drain). This has however the inconvenient that the bulk cannot be bias independently. In that particular case, having the top well contacts close to the device can minimize the bulk resistance. Back contacting 10 Ω.cm wafer easily leads to a kΩ series resistance.

3.4.2 Reduce gate conductance

The gate conductance scales linearly with the test structure area. As presented in Figure 3.28(a), the impact of the gate leakage current on the C-V characteristic is diminishing with scaling the test structure area. However, the area cannot be chosen too small since the instrument precision as well as the uncertainty in the device area will limit the scaling of the test structure.

As shown in Figure 3.28(b), the phase angle is also improving (~90^o phase angle obtained on a large gate bias range) when scaling the test structure area. Note that this does not mean that the instrumentation error is decreasing when scaling the test structure; C and G scaling equally with the test structure area. From the expression of D in (3.4), scaling the test structure area will not change D since both G_m and C_m will scale. However, when considering a more accurate equivalent circuit such as the 3-element equivalent circuit (presented in Figure 3.4(c)), D has an area dependent term as shown in (3.5). D and therefore the instrumentation error are also scaling with the test structure area.

In addition to the impact of the external resistance, the gate bond pad capacitance can play also a role. Because the gate bond pad capacitance becomes more and more important when scaling the area, the test structure looks more and more like an ideal capacitor. This could also explain the phase angle improvement (i.e. increase)..

Nevertheless, we believe that the main component for this effect is the external resistance. Unfortunately, the intrinsic device is just as bad and therefore the measurement will not be more accurate even though the phase shift angle is increasing (dissipation factor is lower).

$$D = Rs.(\frac{G^2}{2\pi fC} + \omega.C) + \frac{G}{2\pi fC} \propto D_{area\_dependent} + \frac{G}{2\pi fC} \qquad (3.5)$$

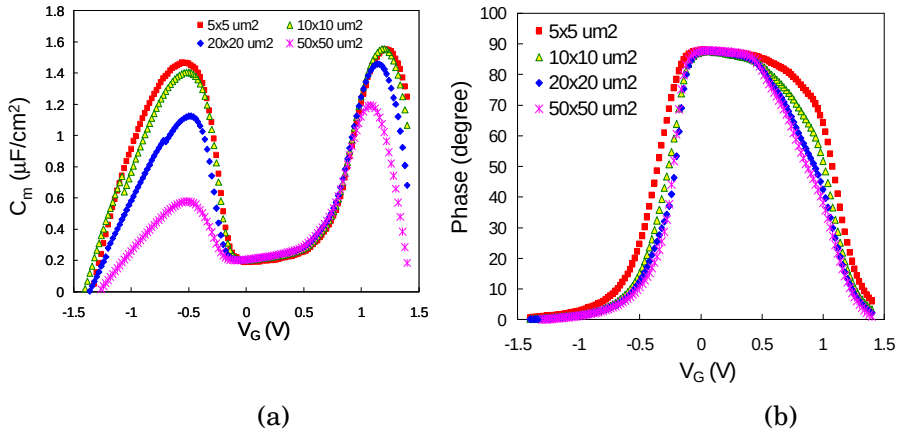


Figure 3.28: *(a) Measured C-V characteristics and (b) its associated phase angles for PMOS gated diodes with various areas (ranging from 2500 down to 25 μm^2) having an ultra-thin oxynitride gate dielectric (J_G=60 A/cm^2 @ V_{GS}= -1 V).*

3.4.3 Control of parasitic elements

The best way to correct for the parasitic elements is to perform a de-embedding of these parasitics on dedicated OPEN and SHORT test structures. These test structures should be design such as the parasitic elements can be easily identify and reproducibly measured.

3.4.4 Design of new test structure

The layout of the C-V test structure is crucial and determines the parasitic resistance, capacitance and inductance. All the considerations mentioned in the above sub-chapters have been taken into account in the design of a new C-V test structure, presented in Figure 3.29. This test structure consists of many gates connected in parallel and connected on both ends to minimize the gate resistance. The amount of polysilicon gate fingers, their width and length should be well chosen to achieve the desired device area while still minimizing the gate resistance (R_G) [22]. R_G is described as:

$$R_G \propto \frac{W}{L} \times \frac{1}{folding\_factor \times N} \tag{3.6}$$

where W and L are the width and length of the polysilicon gate fingers, and N the number of gates.

The well, source and drain areas are shorted to minimize R_S. Care should be taken to choose not too long gate lengths to reduce the channel resistance. To avoid effects of channel resistance in both inversion and accumulation the distances between active/gate and well/gate should be minimized. A large amount of bulk and active (*S*/*D*) contacts are drawn at minimum distance of the device. The gate bondpad used polysilicon (grounded) to shield the bondpad from the substrate. This will strongly reduce the pad capcitance.

Note that this structure can be designed for two-port RF characterization in ground-signal-ground configuration [23]. The gate is connected to Port 1, the source/drain to Port 2 and the well to ground. The gate is biased while other terminals are grounded.

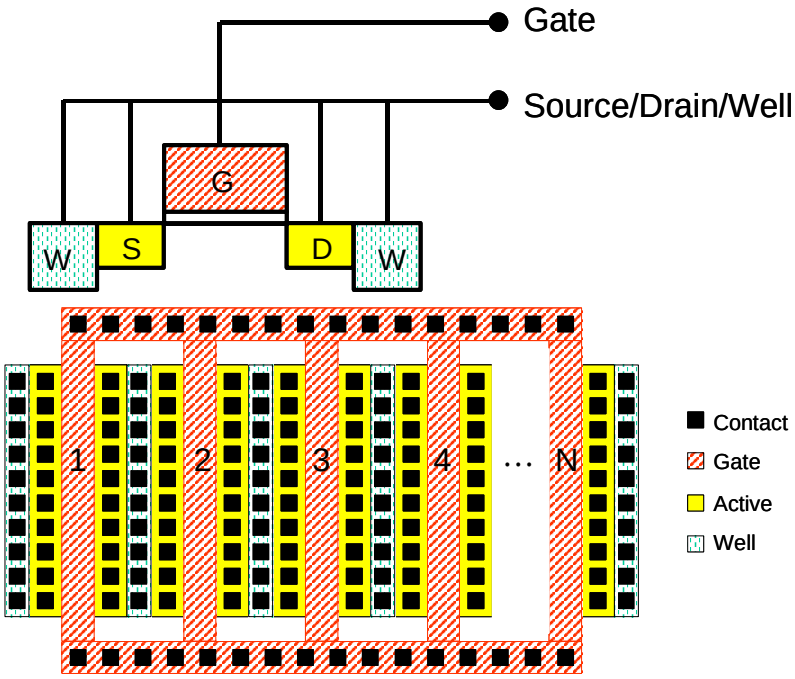


Figure 3.29: *(a) Cross section and (b) layout of a new MOS capacitor test structure. The number of the polysilicon gate fingers (N) should be chosen to both reduce the gate resistance and achieved the desired total capacitor area.*

Dedicated SHORT and OPEN test structures should be designed based on the design of Figure 3.29. The OPEN structure is designed without the active and polysilicon areas and the SHORT one is done at the metal 1 level and on field.
This test structure can be used both for HF and RF C-V measurements.

3.5 Parameter extraction from C-V curves

3.5.1 Introduction

The measurement of C-V curves is targeted towards the determination of key parameters of the MOS structure: oxide thickness, flat band voltage, fixed charge, interface state density, substrate doping concentration and gate depletion. The classical C-V theory [24-30] is deficient in two main areas when applied to ultra-thin gate oxides. Indeed, the theory must include:

- *Polysilicon depletion effects*

Due to the finite doping concentration and insufficient activation of dopants in the polysilicon gate, a depletion layer forms at the polysilicon/gate dielectric interface when applying a gate voltage to bias the transistor in conduction. This will cause a voltage drop in the polysilicon gate [31-32]. The polysilicon depletion will depend both on the doping of the polysilicon gate and on the type of dopants. As a result, the effective oxide thickness is increased (~3-4 Å increase for current CMOS 65 nm technology) .

- *Quantum mechanical surface quantization effects*

Because of the conduction band triangular like well at the surface, electron states are a series of discrete levels above the edge of the conduction band. From a device viewpoint, this has two major effects:

1) For the same semiconductor inversion layer charge, the semiconductor surface potential is larger than classically predicted;
2) Charge is located further from the surface than classically predicted.

As a consequence, both the effective oxide thickness (~4 Å increase) and the depletion layer charge density are increased.

3.5.2 C-V modeling methods

In order to extract MOS key parameters, various methods have been proposed and some of them will be described below. In all cases, the C-V curve should be leakage and parasitic free to allow reliable parameters extraction.
There have been many models used to determine device parameters based on either analytical formulations or numerical calculations of the Poisson and/or Schroedinger equations [33]–[43]. A comparison of some models can be found in the literature

[44]. There is still debate as to the type of model or simulation that is most physically correct.

3.5.2.1 NCSU CVC model

This model uses a non linear least squares curve fitting to the exact C-V equation with adjustable parameters: flat band voltage (V_{FB}), polysilicon doping (N_A) and oxide thickness (T_{ox}) [43]. This method requires to model the semiconductor charge as a function of voltage (one energy subband considered). The model is presently widely used by most of SEMATECH company members for the determination of the T_{ox}, V_{FB}, and substrate doping from the weak accumulation regime. This model is very limited if the C-V curve is not leakage free, as shown in Figure 3.30.

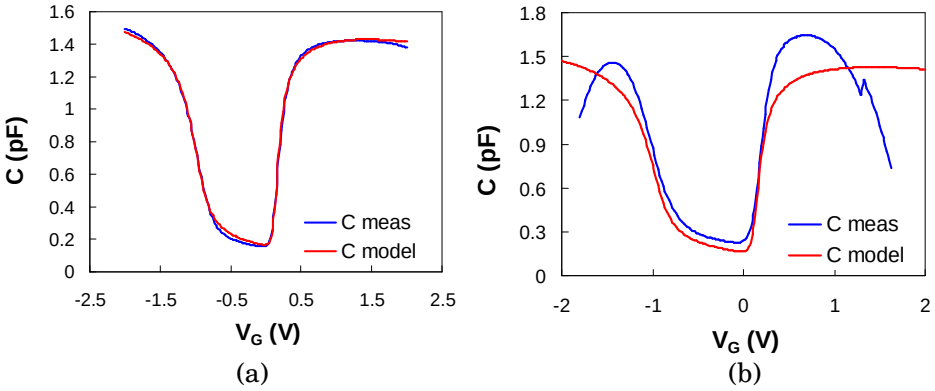


Figure 3.30: *C-V curves measured at 1 MHz on NMOS devices with a (a) thick (J_G < 1 A/cm² at V_{GS}=1 V) and (b) thin (J_G > 1 A/cm² at V_{GS}=1 V) gate dielectrics and the C-V characteristics obtained in fitting the experimental data with the CVC model [42]. A good fit is obtained only if the C-V curve is leakage free.*

3.5.2.2 MOS Model 11

As mentioned in section 3.2.2.2, the compact model MOS model 11 (successor of MOS model 9) is a surface-potential-based model [17]. It allows a full fit of the C-V curve and extraction of T_{ox}, V_{FB}, substrate doping concentration, and gate depletion.

3.5.3 Extraction of relevant parameters: comparison of models

A comparison of the CVC and MOS model 11 (MM11) models has been made on parasitic and external resistances free C-V curves (see chapter 3.3), as presented in Figure 3.31.

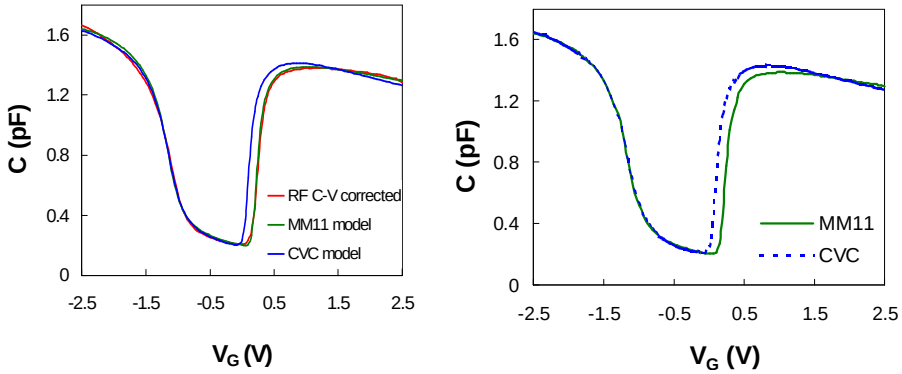


Figure 3.31: *(a) C-V curve corrected for external resistances with its simulated CV curves using MM11 and CVC models. (b) The curve obtained by MM11 and the related fit using the CVC model is also shown. Whereas the CVC model results in a good fit in accumulation a large error is observed, in inversion. Excellent fit was obtained in both regimes (accumulation and inversion) when using MM11. The extracted parameters are reported in table 2.*

A very good fit was obtained when using MM11 while a large error was observed when applying CVC in the depletion to inversion part of the C-V curve. The reason of this discrepancy could be a wrong estimation of the polysilicon doping concentration. However, no improvement was observed when adding the polysilicon doping concentration as a varying parameter in the model. Nevertheless, the two models give reasonable CET_{inv} values as compared to the measured C-V curve, as presented in table 1 ($\Delta CET_{inv} < 0.5$ Å).

	CET_{inv} (Å)	EOT (Å)	N_{poly} ($at.cm^{-2}$)
C-V measured	20	-	-
CVC	19.6	13	1.5×10^{20}
MM11	19.9	14.7	3×10^{20}

(a)

	CET_{inv} (Å)	EOT (Å)	N_{poly} ($at.cm^{-2}$)
CVC	19.3	12.7	1.5×10^{20}
MM11	19.9	14.7	3×10^{20}

(b)

Table 3.1: *Summary of extracted parameters using CVC and MM11 models on (a) measured C-V curve (from Figure 3.31(a)). The polysilicon doping concentration (N_{poly}) is used as an input parameter. The fitting CVC model was also applied to the C-V curve obtained with MM11 (from Figure 3.31(b)) for better comparison of the two models (b). The CET_{inv} parameter is measured at V_{GS}=1 V while the EOT is extracted from the CVC or MM11 models.*

The EOT values extracted from the CVC or MM11 C-V fitted curves as well as the polysilicon doping levels (N_{poly}) used as input parameters for the two models are reported in Table 3.1. A larger polysilicon doping concentration (two times greater) was needed to get a good fit with MM11 as compared to CVC. From SIMS data, it seems that 3×10^{20} at.cm^{-2} is the most accurate value for the n+ polysilicon doping concentration. A non negligible difference in EOT is observed when using the CVC or MM11 models (ΔEOT > 1 Å). This shows that this parameter is very dependent on the model used. Therefore, care should be taken in using the same fitting model when benchmarking the EOT parameter of various gate dielectrics.
It can be concluded that the CET_{inv} is a reliable parameter, assessing the electrical thickness of the gate dielectric, that can either be directly measured from a leakage free C-V curve or from a fitted C-V curve obtained with CVC or MM11 models. However, this parameter is not only sensitive to variations in the physical thickness of the dielectric film but also to changes in the polysilicon gate activation.

3.6 Conclusions and recommendations

3.6.1 Conclusions

We have shown that standard C-V measurements (up to 1 MHz) on devices with highly leaky gate dielectric yield large measurement errors. These errors could be minimized in two ways:
Reducing the test structure area while minimizing the contact and well resistances of the test structure. This solution will however be limited by the instrumentation precision, the parasitic elements and the uncertainty of the size of the device. Parasitic elements such as the pad capacitance can be minimized by adding a third well (triple well type of structure). We believe that a limit will be reached for a gate leakage current density of about 100 A/cm^2.
Increasing the measurement frequency. We have successfully demonstrated C-V measurement at RF frequencies on devices with gate leakage current densities up to 1000 A/cm^2 providing excellent de-embedding and accurate selection of the measurement frequencies have been done. We believe that this method can be extended to even leakier gate dielectrics (although irrelevant for CMOS technologies) by decreasing the test structure area, performing excellent de-embedding and increasing the measurement frequency range. Note that this technique is time consuming and requires specific measurement set-up.
A benchmark of the two C-V measurement techniques described in this report have been done on the same test structure, as presented in Figure 3.32. The device used for this comparison was an NMOS device with two gate fingers of 20 µm width and 1 µm length. This device has an ultra-thin gate dielectric (J_G=100 A/cm^2 at V_{GS}=1 V).

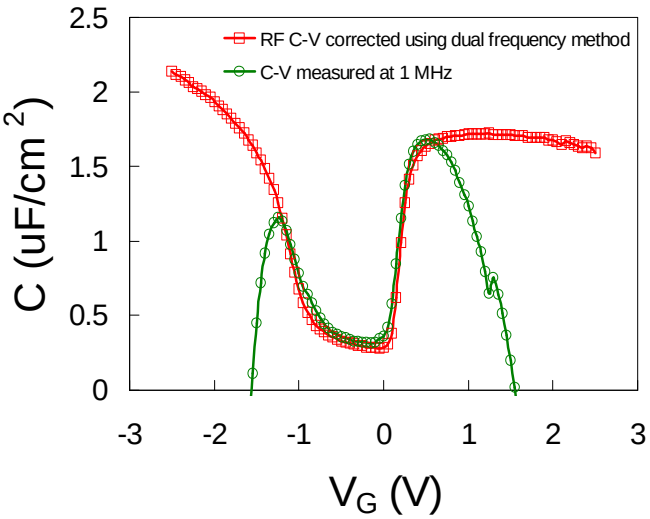


Figure 3.32*: Comparison of C-V curves measured at 1 MHz or at RF frequencies (after applying the dual frequency method [14] to get the intrinsic capacitance).*

It can be observed that, while a strong attenuation is observed on the C-V characteristic measured at 1 MHz, the RF C-V measured curves result in the intrinsic capacitance (after applying dual frequency method [14]).
Finally, a new MOS capacitor test structure is presented that can be used for both HF and RF C-V measurements.

In order to benchmark the thickness parameter of ultra-thin leaky gate dielectrics, care should be taken in having a common:

a) C-V measurement methodology and test structure design resulting in the intrinsic C-V characteristic. We have shown that for non optimized test structure a correction for external resistance is necessary. However, the selection of the accurate MOS equivalent scheme allowing the calculation of the corrected capacitance is not straightforward: there is not one MOS equivalent circuit to be used for any test structure.
b) measurements/extraction methodology for the gate dielectric thickness parameter. For the EOT extracted parameter, same fitting model should be used.

3.7 References

[1] P.A. Kraus, K. Ahmed, T.C. Chua, M. Ershov, H. Karbasi, C.S. Olsen, F. Nouri, J. Holland, R. Zhao, G. Miner and A. Lepert, *Low-energy nitrogen plasmas for 65-nm node oxynitride gate dielectrics: a*

correlation of plasma characteristics and device parameters, Symp. on VLSI Tech., pp. 143-144, 2003.

[2] F.N. Cubaynes, C.J.J. Dachs, C.Detcheverry, A. Zegers, V.C. Venezia, J. Schmitz, P.A. Stolk, M. Jurczak, K. Henson, R. Degraeve, A. Rothshild, T. Conard, J. Petry, M. Da Rold, M. Schaekers, G. Badenes, L. Date, D. Pique, H.N. Al-Shareef, R.W. Murto, *Gate dielectrics for high performance and low power CMOS SoC applications*, ESSDERC Conf. Proceedings, pp. 427-430, 2002.

[3] C.A. Richter, A.R. Hefner and E.M. Vogel, *A comparison of quantum mechanical capacitance-voltage simulators*, IEEE Electron Device Lett., Vol 22, no. 1, pp. 35-37, 2001.

[4] R.F. Pierret, *Modular series on solid state devices*, volume IV: Field Effect Devices, Addison-Wesley, Reading 1990; pp. 47-58.

[5] H. Veendrick, *Deep-submicron CMOS Ics, from basics to ASICs*, Kluwer Bedrijsinformatie, Deventer 1998, pp. 36-45.

[6] M. Schulz, *Electrical characterization of the Si-SiO2 system*, Microelectronic Engineering 40 (1998) pp. 113-130.

[7] E.H. Nicollian and J.R. Brews, *MOS (Metal Oxide Semiconductor) Physics and technology*, Wiley, New-York 1992.

[8] S. Wolf, *Silicon Processing for the VLSI era, Volume 3: The Submicron MOSFET*, Lattice press, Sunset Beach 1995; Chapter 3.

[9] K.S. Krisch, J.D. Brude and L. Manchanda, *Gate capacitance attenuation in MOS devices with thin gate dielectrics*, IEEE Electron Device Lett., Vol. 17, no. 11, pp. 521-524, 1996.

[10] C-H. Choi, J.S. Goo, T-Y Oh, Z. yu, R.W. Dutton, A. Bayoumi, M. Cao, P. Vande Voorde, D. Vook and C.H. Diaz, *MOS C-V characterization of ultrathin gate oxide thickness (1.3-1.8 nm)*, IEEE Electron Device Lett., Vol. 20, no. 6, pp. 292-294, 1999.

[11] E.M. Vogel, W.K. Henson, C.A. Ritcher and J.S. Suehle, *Limitations of conductance to the measurement of the interface state density of*

MOS capacitors with tunneling gate dielectrics, IEEE Trans. Electron Devices, Vol. 47, no. 3, pp. 601-608, 2000.

[12] K.Z. Ahmed, E. Ibok, G.C-F. Yeap, Q. Xiang, B. Ogle, J.J. Wortman and J.R. Hauser, *Impact of tunnel currents and channel resitsance on the characterization of channel inversion layer charge and polysilicon gate depletion of sub-20Å gate oxide MOSFETs*, IEEE Trans. Electron Devices, Vol. 46, no. 8, pp. 1650-1655, 1999.

[13] HP 4284A Operating manual, hewlett-Packard, Japan, 1994.

[14] K.J. Yang and C. hu, *MOS capacitance measurements for high-leakage thin dielectrics*, IEEE Trans. Electron Devices, Vol. 46, no. 7, pp. 1500-1501, 1999.

[15] D. Douglas W. Barlage, James T. O'Keeffe, Jack T. Kavalieros, Michael M. Nguyen and Robert S. Chau, *Inversion MOS capacitance extraction for high-leakage dielectrics using a transmission line equivalent circuit*, IEEE Electron Device Lett., Vol. 21, no. 9, pp. 454-456, 2000.

[16] R. van Langevelde, A.J. Scholten, R. Duffy, F.N. Cubaynes, M.J. Knitel and D.B.M. Klaassen, *Gate current: Modeling, ΔL extraction and impact on RF performance*, IEDM Tech. Dig., 2001, pp. 289-292.

[17] R. van Langevelde, A.J. Scholten and D.B.M. Klaassen, *Mos Model 11, Level 1101*, www.semiconductors.philips.com/philips_models.

[18] K. Ahmed, E. Ibok and J. Hauser, *Capacitor test structures for C-V measurements on COS devices with sub-20Å oxides*, ESSDERC Conf. Proceedings, 2000.

[19] M.A.C.M. Koolen, J.A.M. Geelen and M.P.J.G. Versleijen, *An improved de-embedding technique for on-wafer-high-frequency characterization*, Proceedings BCTM, pp. 188-191, 1991.

[20] J. Schmitz, F.N. Cubaynes, R.J. Havens, R. de Kort, A.J. Scholten and L.F.Tiemeijer, *RF Capacitance-Voltage Characterization of MOSFETs with high leakage dielectrics*, IEEE Electron Device Lett., Vol. 24, no. 1, pp. 37-39, 2003.

[21] J. Schmitz, F. Cubaynes, R de Kort, R. Havens, A. Scholten and L. Tiemeijer, *The RF C-V method for characterization of leaky gate dielectrics*, INFOS conf., 2003.

[22] W. Jeamsaksiri, A. Mercha, J. Ramos, S. Decoutere and F.N. Cubaynes, *Optimal frequency range selection for full C-V characterization above 45 MHz for ultra-thin (1.2 nm) nitrided oxide MOSFETs,* submitted to ICMTS conf., 2004.

[23] P.H. Woerlee and P.W.H. de Vreede, *RF characterisation of 0.25 and 0.18 μm CMOS devices*, NL-Report 7067, Philips Electronics N.V., 1999.

[24] J. Schmitz, F. Cubaynes, R de Kort, R. Havens, and L. Tiemeijer, *Test Structures for C-V Measurements in the Radio Frequency Range*, Proc. ICMTS conf., pp. 181-185, 2003.

[25] J. Maserjian, G. Petersson and C. Svensson, *Saturation capacitance of thin oxide MOS structures and the effective surface density of states of silicon*, Solid State Electron., Vol. 17, pp. 335-339, 1974.

[26] M.J. McNutt and C-T. Sah, *Determination of the MOS oxide capacitance*, J. Appl. Phys., Vol. 46, pp. 3909-3913, 1975

[27] B. Ricco, G.Tondi and M. Lanzoni, *Extraction of oxide thickness from harmonic distortion of displacement currents in MOS capacitors*, IEEE Trans. Electron Devices, Vol. 44, no. 9, pp. 1552-1554, 1997.

[28] B. Majkusiak and A. Jakubowski, *A technical formula for determining the insulator capacitance in a MOS structure*, Solid State Electron.,Vol. 35, pp. 223-224, 1992.

[29] B. Ricco, P.Olivo, T.N. Nguyen, T-S. Kuan and G. Ferriani, *Oxide-thickness determination in thin-insulator MOS structures*, IEEE Trans. Electron Devices, Vol. 35, no. 4, pp. 432-438, 1988.

[30] S. Walstra and C-T. Sah, *Extension of the McNutt-Sah method for measuring thin oxide thicknesses of MOS devices*, Solid State Electron.,Vol. 42, pp. 671-673, 1998.

Chapter 3

[31] H. Reisinger, H. Oppolzer and W. Hönlein, *Thickness determination of thin SiO_2 on silicon*, solid State electron., Vol. 35, pp. 797-803, 1992.

[32] C-Y. Lu, J.M. sung, H.C. Kirsch, S.J. Hillenius, T.E. Smith and L. Manchanda, *Anomalous C-V characetristics of implanted poly MOS structure in n+/p+ dual-gate CMOS technology*, IEEE Electron Device Lett., Vol. 10, pp. 192-194, 1989.

[33] B. Ricco, R. Versari and D. Esseni, *Characterization of polysilicon gate depletion in MOS structure*, IEEE Electron Device Lett., Vol. 17, no. 3, pp. 103-105, 1996.

[34] S.A. Hareland, S. Krishnamurthy, S. Jallepalli, C-F Yeap, K. Hasnat, A.F. Tasch and C.M. Maziar, *A computationally efficient model for inversion layer quantization effects in deep submicron N-channel MOSFETs*, IEEE Trans. Electron Devices, vol. 43, no. 1, p90-96, 1996.

[35] M. J. v. Dort, *A simple model for quantization effects in heavily doped silicon MOSFETs at inversion conditions*, Solid-State Electron., vol. 37, pp. 435, 1994.

[36] R. Rios and N. Arora, *Determination of ultra-thin gate oxide thicknesses for CMOS structures using quantum effects*, IEDM Tech. Dig., 1994, pp. 613-615.

[37] S. Jallepalli, J. Bude, W.-K. Shih, M.R. Pinto, C.M. Maziar and A.F. Tasch, *Electron and hole quantization and their impact on deep submicron silicon p- and n-MOSFET characteristics*, IEEE Trans. Electron Devices, vol. 44, no. 2, pp. 297-303, 1997.

[38] S. A. Hareland, S. Jallepalli, W.-K. Shih, H. Wang, G.L. Chindalore, A.F. Tasch and C. M. Maziar, *A physically-based model for quantization effects in hole inversion layers*, IEEE Trans. Electron Devices, vol. 45, no. 1, pp. 179-186, 1998.

[39] T. Janik and B. Majkusiak, *Analysis of the MOS transistor based on the self-consistent solution to the Schrodinger and poisson equations and on the local mobility model*, IEEE Trans. Electron Devices, vol. 45, no. 6, pp. 1263-1271, 1998.

[40] S. H. Lo, D. A. Buchanan, Y. Taur, andW.Wang, *Quantum-mechanical modeling of electron tunneling current from the inversion layer of ultrathin- oxide nMOSFET's*, IEEE Electron Device Lett., vol. 18, pp. 209-211, 1997.

[41] F. Rana, S. Tiwari, and D. A. Buchanan, *Self-consistent modeling of accumulation layers and tunneling currents through very thin oxides*, Appl. Phys. Lett., vol. 69, pp. 1104–1106, 1996.

[42] K. S. Krisch, J. D. Bude, and L. Manchanda, *Gate capacitance attenuation in MOS devices with thin gate dielectrics*, IEEE Electron Device Lett., vol. 17, pp. 521-524, 1996.

[43] F.P. Widdershoven, *Extraction of gate oxide thickness from C-V measurements*, Proc. INFOS 2001, (2001).

[44] J.R. Hauser and K. Ahmed, *Characterization of ultra-thin oxides using electrical C-V and I-V measurements*, Characterization and Metrology for ULSI technology, Int. Conf. 1998.

[45] G. Tempel, *Assesment of different CV simulation tools*, GSEWG Review meeting, May 22-24 2001.

Chapter 4
Physical characterization of ultra-thin oxide based films

4.1 Introduction

4.1.1 Introduction

Because the amount and distribution of nitrogen in the oxide dielectric film have a direct impact on the device performance, precise and accurate characterization methods are required. The physical characterization of sub-3.0 nm nitrided oxide films represents an incredible challenge due to the ultra-thin physical thickness and the hybrid nature of the film. Conventional material characterization techniques need to be optimized to be applicable to MOS structures with ultra-thin films. There are two essential measurements enabling the optimization of ultra-thin nitrided oxide films:

1. The first one is the thickness measurement. A standard and wide spread thickness measurement is the ellipsometry measurement. However, this technique has a limited resolution. Moreover, ellipsometry measurements are very difficult to interpret for complex dielectrics such as nitrided oxide since the refractive index changes with the composition of the film. Transmission Electron Microscopy (TEM) is in principle a technique which provides the thickness of thin layers in a very direct way. Yet, TEM preparation and TEM analysis are relatively time-consuming; as a consequence it is not a feasible technique for large series of samples. Moreover, the interpretation of TEM cross-section pictures related to the thickness of ultra-thin oxide layers (< 2 nm) is not straightforward, as it will be shown in this chapter.

2. The second one is the characterization of the nitridation process itself: nitrogen (N) content, N distribution and bonding

configuration of the atoms comprised in the film. Lots of analytical techniques are available for physical characterization of dielectric films. Some of them are summarized in Table 4.4.1 where the detection limits, depth resolution, and present limitations are detailed.

Analytical Technique	Signal Detected	Detection Limits	Information depth	Depth Resolution [1]	Limitations
Auger Electron Spectroscopy	Auger electrons from near surface atoms	0.1 - 1 at %	0.5 - 10 nm	< 1 nm	No chemical information; not applicable on non conducting materials
AFM	Atomic scale roughness	---	0.01 nm		No profile capability, no capability to distinguish N from O
STM	Atomic scale roughness	---	0.01 nm		No profile capability, no N / O distinction, and sample must be conducting
FTIR	Infrared absorption	0.1 - 100 ppm	> 10 nm	> 10 nm	Requires thickness sample and improved Signal to noise (~200 nm thick)
XPS	Photoelectrons	0.01 - 1 at %	1 -10 nm	1 nm	Lateral resolution at best 10 μm
Synchr. XPS	Photoelectrons	0.01 - 1 at %	0.3 – 10 nm	1 nm	Lateral resolution < 1 μm
Ellipsometry	Polarization state of electromagnetic radiation	-	0.3 – 1000 nm	0.2 nm	Surface contamination can influence the measurement
SIMS	Secondary ions	1E12 - 1E16 at/cm^3	< 1 nm	< 1 nm	Not as effective on insulators
TOFSIMS	Secondary ions, atoms, molecules	10^{16} - 10^{18} at/cm^3	< 1 nm	≈ 1 nm	Calibration required on very similar reference samples (matrix effects)
TEM	Secondary and backscattered electrons	-	-	0.1 nm	Require extensive sample preparation
RBS	Backscattered	1 - 10 at %	2 - 500 nm	2 - 20 nm	Difficulty in

[1] The best attainable depth resolution in profiling is given

He atoms	(Z<20) 0.01-1 at % (20<Z<70) 0.001-0.01 at % (Z>70)	determining depth-profiles in films < 2.0 nm thick

Table 4.4.1: *Summary table of various analytical techniques detailing the detection limits, depth resolution, and present limitation [46].*

In our study, ellipsometry, Rutherford Backscattering Spectrometry (RBS), TEM, X-ray Photo-electron Spectroscopy (XPS) and Time of Flight Secondary Ion Mass Spectroscopy (TOFSIMS) have been used to fully characterize ultra-thin nitrided oxide films. This choice is first based on the availability of such techniques and on their capabilities to characterize ultra-thin layers. Indeed, with these techniques, information on the concentration, the profile and the chemical bonding configuration of the various elements present in ultra-thin nitrided oxide films are collected as well as the physical thickness of such hybrid layers.
It is important to point out that none of these measurements have been performed without vacuum break between the processing of the ultra-thin films and the measurements. However, care was taken to close couple processing and measurements.
Some of these techniques had to be optimized during this work to enable the characterization of ultra-thin oxide based thin films. These techniques have been benchmarked to each other to assess the most appropriate measurement techniques for such thin films.

4.1.2 Chapter overview

An overview of the various physical analysis techniques (ellipsometry, XPS, RBS, TOFSIMS and TEM) chosen to characterize ultra-thin nitrided oxides is given in part 2. The experimental set-up, optimized for the characterization of ultra-thin oxide based layers, is detailed for each technique.
While the use of pure oxide dielectrics is limited in advanced CMOS technologies (sub-100 nm nodes), an ultra-thin silicon oxide interface layer is always present prior to any growth or deposition of material having a higher permittivity. The accurate determination of the thickness and composition of such ultra-thin oxide films is therefore of first importance. A thorough study of ultra-thin oxide layers (0.14 – 2 nm) is therefore presented in part 3. A benchmark of four analysis techniques, namely ellipsometry, XPS, RBS and TEM, is proposed for the determination of the thickness of such ultra-thin oxide films.
In section 4 of this chapter, the physical properties of ultra-thin plasma nitridation oxides are presented. The N incorporation mechanism, its concentration and depth profile within the layer have been studied using ellipsometry, TOFSIMS, XPS and RBS analysis. A comparison of the N content and profile obtained with these techniques is shown. The thickness of the studied plasma nitrided films have been

also measured using these techniques. For TOFSIMS analysis, a specific concentration and depth scale calibration has been developed to enable the characterization of ultra-thin plasma nitrided films.

4.2 Overview of some characterization techniques

4.2.1 Ellipsometry measurement

Ellipsometry is a very sensitive non-destructive measurement technique that uses polarized light to characterize thin films, surfaces and material microstructure. It derives its sensitivity from the determination of the relative phase change in a beam of reflected polarized light. Any physical effect which induces changes in a material's optical properties can be studied with such technique. In our work, ellipsometry measurements are used to measure the thickness of ultra-thin gate dielectrics, their uniformity, roughness as well as their stoichiometry (e.g. gate dielectric stack).
Based on ellipsometry measurements, a method has been developed to estimate the amount of N and the N profile within ultra-thin nitrided oxide films has been developed. This is presented in the last part of section 2.1.

4.2.1.1 Theory

Ellipsometry involves the reflection of light from a surface [47], [48]. Figure 4.1 illustrates the basic principle behind ellipsometry. First, the polarization state of incoming light is known. Plane polarized waves that are in the plane of incidence are referred as *p*-waves and plane polarized waves perpendicular to the plane of incidence as *s*-waves. Then the incident light interacts with the sample and reflects from it. The interaction of the light with the sample causes a polarization change in the light, from linear to elliptical polarization. The polarization changes, or a change in the shape of the polarization, is then measured by analyzing the light reflected from the sample.

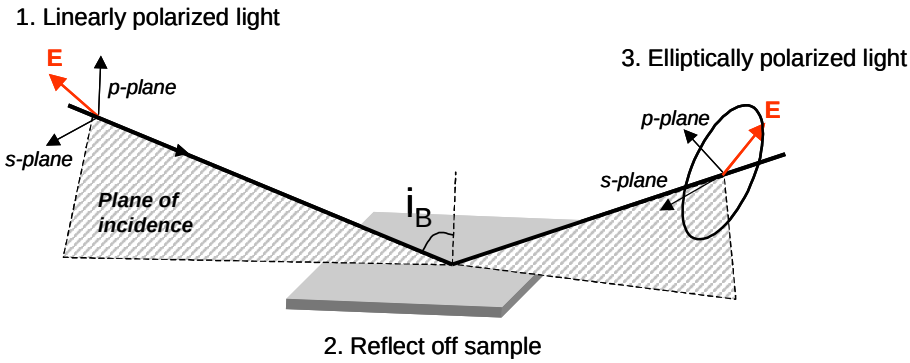


Figure 4.1: *Geometry of an Ellipsometry Measurement. The ellipsometer operates with illumination at angle of incidence near the Brewster's angle for best sensitivity (~76°).*

Ellipsometry measures two values, Psi (Ψ) and Delta (Δ), that describe this polarization change. These values are related to the ratio of Fresnel reflection coefficients, *Rp* and *Rs* for *p*- and *s*- polarized light, respectively. *Rp* (or *Rs*) is the ratio of the amplitude of the reflected wave to the amplitude of the incident wave for a single interface.

$$\tan\Psi e^{i\Delta} = \frac{R_p}{R_s} \tag{4.1}$$

where $\tan\Psi$ is the intensity ratio of the two components, and Δ is the phase difference between them.

Because ellipsometry measures the ratio of two values, it can be highly accurate and very reproducible. The ratio is a complex number, thus it contains "phase" information, Δ, which makes the measurement very sensitive and applicable for ultra-thin films.

Ellipsometry does not measure optical constants or film thickness, however Ψ and Δ are functions of these characteristics. Film thickness and optical constants are thus extracted through a model based analysis using optical physics (Fresnel reflection coefficients, Snell's law, etc...) [48], [49].

4.2.1.2 Measurement set-up

Typical commercial ellipsometer systems obtain the Ψ and Δ signals indirectly through transformations of measured intensity signals, and, unlike a reflectometer, there is no need for an absolute reference. Since both the intensity response Ψ and the phase response Δ are measured, more information about the sample can be extracted from these signals.

The rotative polarizer configuration has been used by many state-of-the-art commercial ellipsometer systems, such as SOPRA's GESP5 and KLA-Tencor's F5. The basic optical configuration is illustrated in Figure 4.2. The optical path consists

of the broadband light source, two rotatable polarizing filters known as the polarizer and the analyzer, the sample, and the spectrometer. During the measurement, the analyzer stays at a certain position and the polarizer rotates continuously to create time-variant signal at the spectrometer.

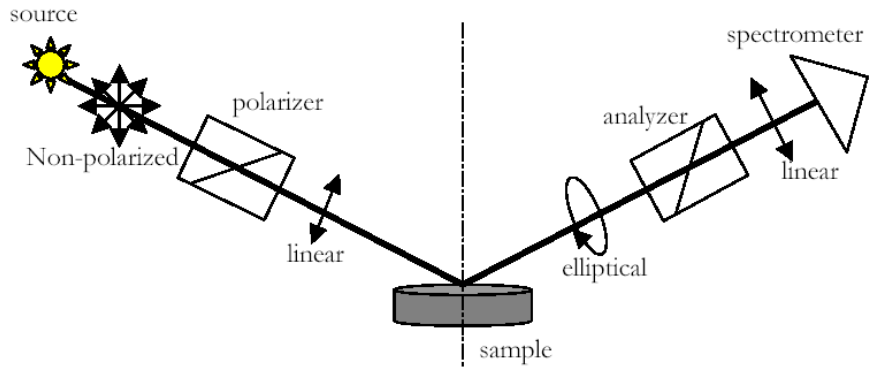


Figure 4.2: *Illustration of rotating-polarizer ellipsometer setup*

The advantage of the rotating polarizer technique is that it is optically and mechanically simple. Only polarizers and focusing lenses (focusing reflecting mirrors in production configuration) are used in the light path, and these optical elements are relatively easy to make and characterize.
In the work performed during this thesis, the KLA-Tencor F5 was used for ellipsometry measurements.

4.2.1.3 Estimation of N content and profile in ultra-thin nitrided oxide layers using ellipsometry measurement

An in-line technique called D2R (Delay To Reoxidation) was developed in order to monitor the concentration of interfacial N atoms. It is based on the retardation of the oxide growth after N incorporation. In Figure 4.3, a schematic description of the measurement procedure is shown: reference bare Si and studied wafers are first pre-oxidized using a similar process. Subsequently to the nitridation of the only studied sample, the two wafers are re-oxidized using a 20 nm dry oxidation (on silicon) recipe. Resulting optical thickness are finally compared and D2R is computed as:

$$D2R = \frac{T_4 - T_3}{T_4 - T_2} \times 100 \qquad (4.2)$$

where T_2, T_3 are the optical oxide thickness after nitridation and after reoxidation of the nitrided oxide, respectively. T_4 is the optical oxide thickness after the reoxidation treatment on a bare Si wafer.

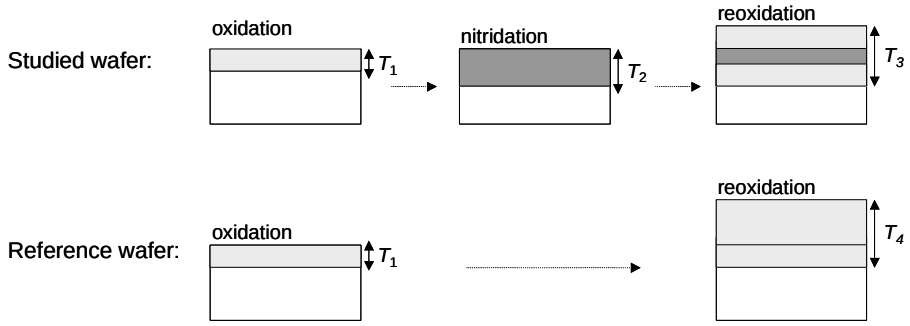


Figure 4.3: *Description of the process steps and corresponding optical thickness measurements necessary to compute the so-called D2R parameter.*

Note that 0 % D2R refers to a pure oxide interface with silicon, 100 % D2R refers to a pure nitride interface with silicon. We will see later in this chapter that for ultra-thin plasma nitrided oxides, the D2R parameter is in very good agreement with the total nitrogen concentration in the film.

4.2.2 X-ray Photoelectron Spectroscopy (XPS)

X-ray Photoelectron Spectroscopy (XPS) is a non-destructive technique that allows to study the composition of thin films by studying binding energies. It is also used to measure the thickness of ultra-thin films and has the advantage not to be sensitive to contamination layers at the surface, contrary to ellipsometry measurement techniques.

4.2.2.1 Definition

In this technique the surface of a sample is irradiated with soft X-rays [50]. An interaction of the X-rays with the atoms causes the emission of photoelectrons, known as the photoelectric effect. The measured kinetic energy of the emitted photoelectrons is given by:

$$K_e = h\upsilon - B_e - \phi \tag{4.3}$$

where $h\upsilon$ is the energy of the X-ray photon, B_e is the binding energy of the atomic orbital from which the photoelectron originates and Φ is the work function. The work function is the minimum amount of energy an individual electron has to gain to escape from a particular surface.

For each and every element, there will be a characteristic binding energy associated with each core atomic orbital, i.e. each element will give rise to a characteristic set of peaks in the photoelectron spectrum at kinetic energies determined by the photon energy and the respective binding energies. The presence of peaks at particular energies therefore indicates the presence of a specific element in the sample under study. Furthermore, the intensity of the peaks is related to the concentration of the element within the sampled region. In Figure 4.4, an example of a photoelectron

spectrum from an ultra-thin SiON film on Si is shown. The O1s, N1s and Si2p peaks are of primary importance in the study of SiON layers. The Si2p peaks are the Si as elementary Si (set at 99 eV for calibration) and the Si as SiON peak.
The exact binding energy of an electron depends not only upon the level from which photoemission is occurring, but also upon :

1. the formal oxidation state of the atom
2. the local chemical and physical environment

Changes in either {1} or {2} give rise to small shifts in the peak positions in the spectrum, so-called chemical shifts. Such shifts are readily observable and interpretable in XP spectra (unlike in Auger spectra) because the technique :

- is of high intrinsic resolution (as core levels are discrete and generally of a well-defined energy)
- is a one electron process (thus simplifying the interpretation)

Atoms of a higher positive oxidation state exhibit a higher binding energy due to the extra coulombic interaction between the photo-emitted electron and the ion core. This ability to discriminate between different oxidation states and chemical environments is one of the major strengths of the XPS technique and is thus very powerful in the research of thin hybrid films such as SiON layers.
Before the photoelectrons (extracted from atoms) reach the surface, they travel a certain distance through the matter. This distance depends on several parameters from which two are of great influence: the initial electron energy and the nature of the medium interacting with this electron. While the electron is traveling through matter it can undergo inelastic collisions. These are collisions with (target) electrons in which energy is transferred from the photoelectron, hence the term inelastic. The average distance covered by an electron between two inelastic collisions is called "Inelastic Mean Free Path" (IMFP). This distance is noted as λ. When the electron undergoes a collision, its energy decreases randomly and it contributes to the background noise build up, at binding energies higher than the peak energy. The photoelectrons that leave the matter without an inelastic collision form the peaks. These peaks are asymmetrical for pure metals due to coupling with conduction electrons and symmetrical for insulators. For semiconductors the peak maybe slightly asymmetrical, which is only visible when the measurements are done with a high energy resolution. The observed peak width is a convolution of the natural line width, the width of the X-ray line that created the photoelectron line, and the instrumental resolution.

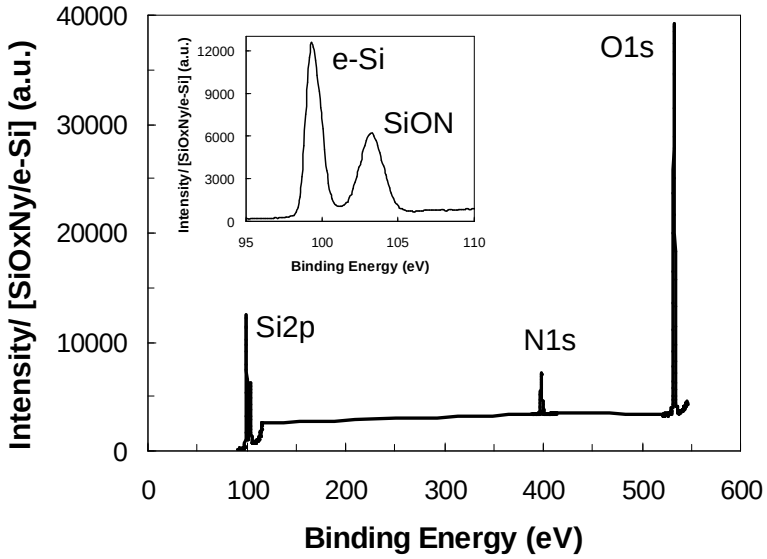


Figure 4.4: *Survey of photoelectron spectrum of SiON on Si. The Si2p peaks are detailed in the inset of the figure. Four peaks are measured: the elementary Si, the Si as SiON, the N1s and O1s peaks.*

The challenge of techniques based on energy probes is the penetration depth that is often many times greater than the film thickness, resulting in degraded signal to noise ratio. The use of a variable energy synchrotron XPS is also an attractive option, from a technological prospective because of its high intensity and relatively low photon energy (typically a few 100 eV and consequently a shorter probe depth than "standard" XPS). However, few synchrotron sources are available which makes it less attractive for the primary technique of dielectric development work in the semiconductor industry.

4.2.2.2 Experimental set-up description

The measurements have been carried out in a Quantum 2000 from Phi [51]. The measurements have been performed using monochromatic AlKα radiation. The measuring spot, pass energy, step-size, entrance angle of the analyzer and the angle between the surface normal of the sample and the analyzer axis (θ) differ for different measurements. The entrance angle of the analyzer was ± 4º if the analyzer-aperture was closed and ± 20º when it was open. For the high resolution measurements the pass energy was set to 11.75 eV and the step-size to 0.025 eV. Most of the measurements were performed with θ = 34º and during all the measurements the analyzer azimuth angle was 22.5º ± 2º. When this geometry is used diffraction effects are minimized [52].

4.2.3 Rutherford Backscattering Spectrometry (RBS)

4.2.3.1 Definition

Bøgh first proposed the application of ion scattering as a surface probe in 1965 [53]. There are numerous experimental approaches to ion scattering that have been developed since then, from which Rutherford Backscattering Spectrometry (RBS) has prevailed as a particularly useful technique for surface and thin film analysis. This technique can analyze the elemental composition as a function of depth in the near-surface region by taking advantage of the well-understood Rutherford scattering cross section for an energetic ion beam interacting with atoms in a solid under pure electrical repulsion between the ion and atomic nuclei [54], [55].

In RBS, a beam monoenergetic ions (H^+ or He^+) is directed at the sample to be investigated. When an energetic ion penetrates the material, it loses energy mainly in collisions with electrons and only occasionally with nuclei. When the positively charged monoenergetic ion comes close to the nucleus of an atom, it will be repelled by the positively charged nucleus. The repulsion force is increasing with the mass of the target atom. The higher the mass of an atom that is hit by a monoenergetic ion, the higher the energy of the ion will be after backscattering (comparable to collisions between billiard balls). This results in mass discrimination. By measuring the energy spectrum of the recoiled ions, information on the composition of the elements, and their depth within the sample can be obtained.

For a surface analysis by RBS, conditions must be such as the mass of surface atoms is considerably higher than the mass of substrate atoms if the peak from surface atoms is to be completely resolved. This is not the case for the study of thin SiO_2 or SiON films on top of a Si substrate, where O and N atoms are lighter than Si atoms. A solution to this problem is to use the channeling effect to make the targeted atoms more or less visible to the projectile monoenergetic ions. Indeed, channeling occurs when the ion beam is carefully aligned with a major symmetry direction of the Si substrate. Most of the ion beam is then steered through the channels formed by the strings of Si atoms. Channeled particle cannot get close enough to the atomic nuclei to undergo large Rutherford scattering, hence scattering is drastically reduced (by a factor of approximately 100). However, atoms such as O and N at the surface produce a peak in the back scattering spectrum because the ion beam is scattered from the surface atoms with the same intensity as from a random array of Si atoms. Therefore, channeling drastically improves the sensitivity of RBS atoms at the surface, enabling the study of ultra-thin layers.

4.2.3.2 Experimental set-up description

With use of a single ended v.d. Graaff accelerator of High Voltage Engineering corporation, a 2 MeV He^+ ion beam is generated. The beam is directed and focused with use of magnets, so the beam impinges the sample along the surface normal, with a small spot size. Channeling in the [100] direction is used in order to suppress the silicon signal under the oxygen or/and nitrogen peak(s). In this way we obtained reasonable statistical accuracies. For the channeling, it is assumed that the channeling direction is perpendicular to the surface, the surface of the silicon wafers being [100]

planes. The used scattering angle, the angle between the detector and the surface normal, was 86.5˚, as depicted in Figure 4.5. The Si surface peak contribution to the total Si peak in the spectrum was used to normalize the integrated charge of He^+ ions. The used surface peak content for [100] silicon is 15.5×10^{15} Si atoms/cm^2. By doing such normalization, the uncertainties in the scattering geometry are eliminated.
An example of an energy spectrum of ions scattered from a sample having a 2.5 nm SiO_2 on top of a Si substrate is presented in Figure 4.6. Accurate calculation of atomic concentrations from the observed scattering yields is performed using a commercial analysis package called RUMP [56].

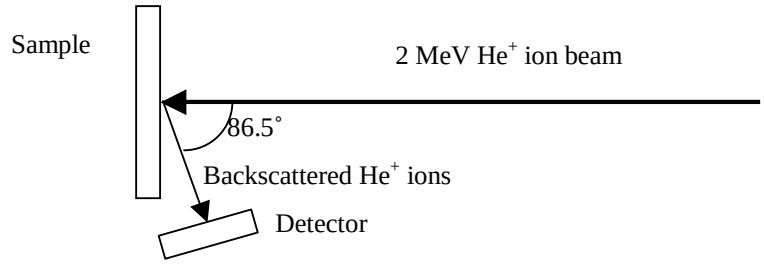


Figure 4.5: *Scattering geometry in RBS.*

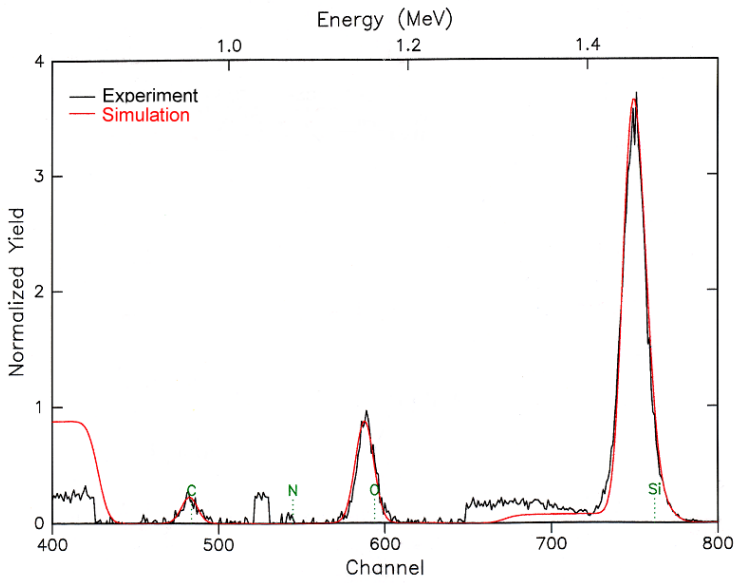


Figure 4.6: *RBS spectrum, measured on a sample having a 2.5 nm SiO_2 on Si. The red curve is the simulation of the spectrum obtained with the RUMP program [11].*

4.2.4 Time of Flight Secondary Ion Mass Spectroscopy (TOFSIMS)

Secondary Ion Mass Spectroscopy (SIMS) is the most widely used technique to characterize the nitrogen profile, although this characterization is extremely difficult. Firstly, ultimate depth resolution (< 1 nm) is required for characterizing ultra-thin layers and thus the lowest possible primary ion energies and/or maximum glancing angles of incidence must be used. Secondly, accurate quantification of the depth profiles is complicated by changes in sputter rate, sputter yield and ionization probability as a function of oxygen and nitrogen content.

4.2.4.1 Definition

Typically, to depth profile the ultrathin oxynitride gate dielectric films, dynamic SIMS uses an Ar^+ (or Cs^+) ion beam and tracks MAr^+ molecular secondary ions to reduce matrix effects and to improve depth resolution, where M represents a matrix species [57], [58]. ArN^+, ArO^+, and $ArSi^+$ secondary ion intensities are examined at low (unit) mass resolution during oxynitride depth profile analysis, using a quadrupole or a double-focused electrostatic sector energy/magnetic sector mass analyzer. However, a 1000 eV beam energy exhibits an Ar^+ (or Cs^+) pre-equilibrium depth of ~ 1.0 nm. Since quantification is not reliable in the pre-equilibrium zone, where the implanted Ar concentration varies with time, the Ar^+ pre-equilibrium zone limits reliable quantification to depths greater than 1.0 nm, which is a substantial fraction of the ultra-thin film thickness [59].

Time-of-flight (TOF) SIMS is a method for depth profiling inorganic oxide films that permits reliable quantification from the uppermost monolayer. This method involves cationization of neutral molecular and elemental species by secondary major matrix cations, instead of primary Ar^+ ions, to track concentrations of various species as a function of depth. Since Me^+ production from a Me_mO_n inorganic oxide substrate is profound, secondary Me^+ ions are a convenient, *in situ* source of cations that can be used from the very outset of the depth profile analysis. Using the Me^+ ions avoids the pre-equilibrium depth zone required to establish an equilibrium concentration of surface species from an external source of primary cations. Recent Monte Carlo simulation studies of sputtered particles from metallic materials indicate that Me_2^+ dimers are formed by a recombination process between independently sputtered Me and Me^+ particles, equivalent to the formation process of the well-known ArM^+ or ArX^+ clusters. Accordingly, molecular matrix cations, MeX^+, are used to track the concentrations of neutral molecular and elemental species, X [60], [61], [62]. Chemically inert ion probe and sputter beams that do not enhance positive or negative ion yields, such as noble gas or Ga sources, are used. As a result, the work function of the surface does not change significantly with depth.

The secondary Si^+ ion peak intensity associated with a silicon oxynitride film is approximately two orders of magnitude more intense than all molecular ion peaks, so the concentration of species X in a silicon dioxide matrix, {X}, is proportional to the SiX^+ molecular ion intensity. Accordingly, Si, O, and N concentration are correlated with Si_2^+, Si_2O^+, and Si_2N^+ molecular ion intensities, since Si-O and Si-N neutral

dimers are simple fragments of the silicate matrix bonding structure. Si_2O^+ and Si_2N^+ intensities are referenced to or normalized by the Si_2^+ matrix ion intensity.
Several depth profiles quantification schemes have been published for different measurement conditions, but they are usually based on comparison with a single standard: typically SiO_2 on Si with some N at the interface with thickness obtained from ellipsometry measurements and N dose from RBS measurements, for example [63], [64], [65]. Characterization of significantly different samples, in particular with higher N content and smaller thickness, by such a method cannot be trusted without further investigations. It is of first importance to determine the erosion rates and sensitivity factors for O and N for various layers. Both the concentration and depth scale have been carefully calibrated for thin SiON films. This work is described in details in section 4.4.1 and in [66].

4.2.4.2 Experimental set-up description

TOFSIMS depth profiling has been performed with dual beam in interlaced mode (alternating sputtering and analysis) on an IONTOF IV instrument using Time-of-Flight detection of secondary ions. The instrument is equipped with a high-current sputter gun and a 3-Lens Ga gun. Details are given below:

	Negative mode
Sputtering	
Primary ion (PI)	200 keV Ar^+
Ion current density	10 nA (300 x 300 μ^2)
Analysis	
Primary ion (analysis)	12 keV Ga^+
Ion current density	0.8 pA (75 x 75 μ^2)
Flooding	-
Mass Resolution	$M/\Delta M$= 8000
Charge compensation	Electron flood gun

Table 4.4.2: *Details of TOFSIMS experimental conditions.*

4.2.5 Transmission Electron Microscopy (TEM)

4.2.5.1 Definition

Transmission Electron Microscopy (TEM) is an analytical technique that is used to study the structure and morphology of materials down to sub nanometer scale. This technique can also be used for layer thickness measurement.
The design of a TEM is analogous to that of an optical microscope. In a TEM, high-energy (>100 keV) electrons are used instead of photons and electromagnetic lenses instead of glass lenses. The electron beam passes an electron transparent sample and an enlarged image is formed using a set of lenses. This image is projected onto a fluorescent screen, a photographic plate or a CCD-camera. Whereas the use of visible

light limits the lateral resolution in an optical microscope to a few tenths of a micrometer, the much smaller wavelength of electrons allows for a resolution of few Angstroms in a TEM.
Image contrast is obtained by interaction of the electron beam with the sample. Several contrast effects play a role. Scattering of the electrons with the sample causes both areas with higher density and areas containing heavier elements to appear darker in the resulting TEM image. Additionally, scattering from crystal planes introduces diffraction contrast. This contrast depends on the orientation of a crystalline area in the sample with respect to the direction of the incoming electron beam. Thus tilting a crystal in the microscope will cause the gray-level in the TEM image of this crystal to change. As a result, each crystal will have its own gray-level in a TEM image of a sample consisting of randomly orientated crystals. In this way one can distinguish between different materials, as well as image individual crystals and crystal defects.

4.2.5.2 Experimental set-up description

The TEM samples are prepared by tripod polishing, i.e. mechanical polishing down to electron transparency. In order to protect the surface of the TEM sample during the preparation, a capping layer needs to be used. Usually, glue in combination with a glass cover is used for capping. The problem with glue as a capping layer is the fact that in a high resolution image it might be difficult to precisely define the interface between the amorphous oxide and the amorphous glue (the glass cover is totally polished away during preparation). For this reason a crystalline capping layer is used, in this case Aluminium (Al). Al is preferred as a capping layer for several reasons. The interface of polycrystalline Al with the amorphous SiO_2 can be imaged very sharply. Glue and other amorphous materials often have a poorly distinguishable interface with oxide layers. Furthermore, Al has approximately the same polishing behavior as SiO_2 and is thus preferred over metals such as W or Pt that have a relatively low polishing rate. The Al was evaporated at low temperature in order to avoid alterations to the oxide layer. In some cases additional Argon ion-milling is used to further thin the TEM sample after finishing the mechanical polishing. TEM studies were performed using a TECNAI F30ST TEM operated at 300 kV.

4.2.5.3 Thickness measurement calibration

In order to ensure accurate as well as reproducible feature measurements, several procedures were followed that will be discussed separately below [67]. All TEM images are taken using the TECNAI F30 ST, operated at 300 kV. The same set of microscope settings and alignments was used for all images. Energy Filtered TEM (EFTEM) was applied; only information from the zero loss peak was used to form the high resolution image (i.e. the image was formed merely using electrons that did not loose any energy to the sample). Zero loss imaging was used as it improves the lateral resolution by removing most of the chromatic aberration. The magnification was calibrated on every image studied. This was performed by taking the Fourier transform (FFT) of the high resolution image of the Si lattice. The periodicity perpendicular to the sample surface was then calibrated using literature data. This direction was chosen for calibration, as it corresponds to the direction in which the

oxide layer thickness is measured. After calibration, an inverse FFT was generated. The resulting calibrated high resolution TEM image was used for accurate determination of the oxide layer thickness. The total error margin for the calibration procedure was estimated to be 0.5%.

An example of a TEM image of ultra-thin SiO_2 film is presented in Figure 4.7. The thickness of the SiO_2 layer was determined using a box, drawn with its edges parallel and perpendicular to the surface normal. Within this box, two parallel lines are drawn that are manually aligned to the SiO_2/Si and SiO_2/Al interfaces. The choice where to situate the interface (i.e. on a row of atoms or between two rows of atoms) results in a 0.10 to 0.15 nm shift in the resulting thickness values. This is the main source for errors. Consequently, we estimate the absolute accuracy of the TEM-thickness values to be ± 0.1 nm.

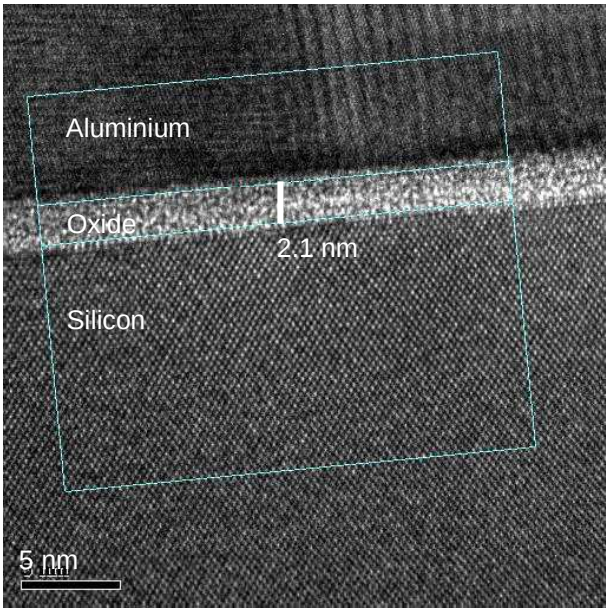


Figure 4.7: *TEM image of an ultra-thin SiO_2 film. The upper layer in the photo shows the aluminum capping-layer. The layer below the aluminum layer is the SiO_2 layer. The box in the picture is used to determine the thickness of the SiO_2 layer. The lattice distance of the elementary silicon was used for the calibration of the TEM photos.*

4.3 Study of ultra-thin silicon oxide film

While the use of "pure" oxide films as the gate dielectric is limited in sub-100 nm CMOS technologies, ultra-thin oxide (SiO_2) layers are very often employed as the interface layers with the silicon (Si) substrate, in combination with a second material grown or deposited on top of it to form the gate dielectric.
As presented in section 2.2, XPS is one of the techniques that is used frequently to determine the thickness and composition of thin layers of SiO_2 on Si. In a recent set of papers, the precise quantification of XPS analyses on thin layers of SiO_2 has been investigated in a systematic way [68], [69]. A comparison between various analysis techniques to determine the thickness and compositions of thin layers of SiO_2 on Si is given in [70] but for SiO_2 layers thicker than 2 nm. Recently, a much more detailed comparison between various analysis techniques for layer thicknesses above 1.5 nm SiO_2 was made by Seah *et al* [71].
For present innovations in the semiconductor industry, SiO_2 layers on Si in the range between 0.14 and 2 nm are more interesting. This motivated us, to start a comparison of SiO_2/Si samples within this range, using four analysis techniques: ellipsometry, XPS, RBS and TEM. The purpose of this work is to determine the optimal way to apply these techniques on ultra-thin SiO_2 films, both experimentally and with respect to the analysis of the results [72].

4.3.1 Experimental

The samples were based upon pure silicon (Si) [100]. After cleaning using an HF last process, a thin oxide film was grown using in-situ steam generation oxidation (ISSG). The thickness of the SiO_2 layer as measured by ellipsometry, $T_{SiO_2}^{opt}$, ranges from 0.14 (wafer measured just after an HF last clean, no oxidation performed) to 3.2 nm. For reference purposes, measurements were also carried out on a "thick" SiO_2 layer on Si: 120 nm, thermally grown. In Table 4.4.3, the optical thickness of the studied SiO_2 films are summarized.
Two series of XPS-measurements were done:

- The first series of measurements was performed with a spot size of 100 µm, a pass energy of 11.75 eV and a step-size of 0.025 eV; the entrance angle of the analyzer was ± 20° unless stated otherwise. A measuring angle Θ of 34° was used and the samples were mounted such that the analyzer azimuth angle was 22.5 ± 2° with respect to the [011] direction. By doing so, the influence of the crystal structure of the substrate upon the results is minimized [68].
- The second series of measurements was performed such that the results could be analyzed with Quases-Tougaard [73]. Extended O1s peaks have been measured with a spot size of 1200 x 500 µm (High Power Mode), a pass energy of 117 eV, a step-size of 0.25 eV and the analyzer entrance angle set at ± 20°. Three values of the measuring angle were used: Θ = 45°, 34° and 0°.

RBS spectra have been recorded and TEM analyses have been made on the same samples. The measurement conditions used for these two techniques are described in sections 2.3 and 2.5, respectively.

Wafer label	Optical thickness (nm)
A	0.137 ± 0.01
B	0.498 ± 0.01
C	1.001 ± 0.015
D	1.007 ± 0.01
E	1.065 ± 0.01
F	1.408 ± 0.035
G	1.421 ± 0.029
H	1.998 ± 0.03
I	2.203 ± 0.073
J	2.51 ± 0.03
K	3.2 ± 0.02
L	120

Table 4.4.3: *Description of the studied SiO_2 films with their optical thicknesses, as determined by ellipsometry measurements.*

4.3.2 Results and discussions

First we consider the results of the XPS measurements. A typical Si2p spectrum is shown in Figure 4.8(a). To determine the peak areas corresponding to elementary Si and the (sub)-oxides of Si, the Si2p peaks were decomposed as follows (using the software package CasaXPS [74]). A Shirley background was subtracted [75]. The best fit for elementary Si (Figure 4.8(b)), as determined from the measurement on sample A (0.14 nm SiO_2), was obtained with two GL(67)T(1.45) curves, a doublet distance of 0.61 eV and a ratio of 2:1. The decomposition into (sub)-oxides was based upon the findings of [76]: doublets of GL(20) curves with a doublet distance of 0.61 eV, equal widths and at 0.97, 1.80, 2.60 and 3.9 +/- 0.2 eV distance from e-Si2p3.
The first approach to determine the thickness of the SiO_2 layer was the use of the standard equation [68]:

$$T_{SiO_2{}^{std}} = L_{SiO_2}(E_{Si})\cos(\Theta)\ln(1+\frac{R_{\exp}}{R_0}) \qquad (4.4)$$

with $T_{SiO2}{}^{std}$ is the SiO_2 thickness obtained from equation (3), $L_{SiO2}(E_{Si})$ the attenuation length for Si2p electrons in SiO_2, R_{exp} the experimental ratio $I_{SiO2}/I_{e\text{-}Si}$ and $R_0=I_{SiO2,\infty}/I_{e\text{-}Si,\infty}$. The parameter $I_{SiO2,\infty}$ denotes the Si2p intensity of "infinitely" thick SiO_2, while $I_{e\text{-}Si,\infty}$ corresponds to the Si2p intensity of pure elementary silicon. We

adopted $L_{SiO2}(E_{Si})$ = 3.448 nm [68]; the measurements have been carried out for Θ = 34°.

The value of R_0 is expected to depend upon the entrance angle of the analyzer. The reason is, that both $I_{SiO2,\infty}$ and $I_{e\text{-}Si,\infty}$ depend upon the entrance angle, but due to the crystal effects in e-Si the dependence of these quantities upon the entrance angle is not identical (see figure 5(b) in [68]). The experimental value of R_0 in our equipment has been determined by measuring a sample of pure Silicon and a sample of "infinitely" thick SiO_2. In the spectrum of the Si2p peak of sample A, no contribution of SiO_2 was detectable (see Figure 4.8(b)). Therefore this sample was considered to be pure silicon. The experimental value of $I_{SiO2,\infty}$ was determined using sample L. The experimental values for $I_{SiO2,\infty}$ and $I_{e\text{-}Si,\infty}$ were corrected for the attenuation of the signals due to a small amount of contamination with hydrocarbons. Combining the experimental values provides for R_0 in the Quantum 2000 at standard conditions (an entrance angle of ± 20°):

$$R_0 = 0.81 \pm 0.02 \qquad (4.5)$$

The uncertainty in the value is due to the background subtraction. Our present value nicely fits into the range of values for R_0 that is found in the literature: values between 0.6 and 0.9 have been reported. We notice, that the experimental value for R_0 obtained in our equipment when a small entrance angle is used (entrance angle ± 4°) is 0.91 ± 0.02. Clearly, R_0 is not a material quantity but rather depends upon the details of the equipment. This is probably one of the reasons for the large variety of values for R_0 found in the literature.

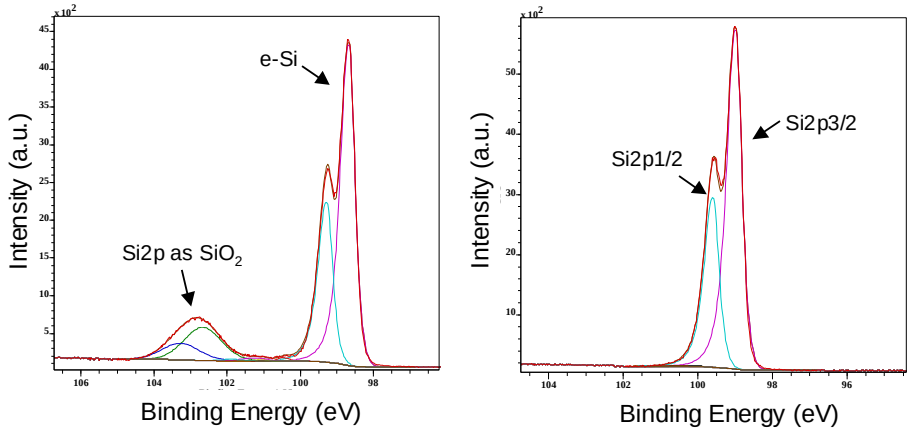


Figure 4.8(a) *Typical example of a Si2p peak. The spin-orbit splitting in the right-hand peak, corresponding to elementary Si, is clearly visible; the FWHM of the components of this peak is 0.31 eV. Sub-oxides are barely present between the peaks of e-Si and SiO_2. (b) Si2p-peak measured for the sample A (optical thickness is 0.14 nm); no SiO_2 contribution is visible.*

Using the standard equation (4.4), we have calculated values for T_{SiO2}^{std} for the series of studied samples. the results can be found in Table 4.4.4. The contribution of sub-oxides was taken into account by adding a weighted average:

$$R_{exp} = \frac{I_{SiO_2} + 0.75I_{Si_2O_3} + 0.5I_{SiO} + 0.25I_{Si_2O}}{I_{e-Si}} \qquad (4.6)$$

The thickness of SiO_2 layers on Si can also be determined by means of an analysis of the e^- loss phenomena of the O1s peak, using the method called Quases-Tougaard [73]. An interesting property of this method is, that it is independent of crystal effects, because the amount of oxygen is being determined. Measurements of the extended O1s peaks have been carried out for Θ = 0°, 34° and 45°. Extended SiO_2 peaks of sample L were used as a reference (to "scale" the spectra). Average values of the thickness of the SiO_2 layers (T_{SiO2}^{QT}) obtained with this method are given in Table 4.4.4.

Finally, the XPS results were also analyzed using a model calculation, in which the samples are assumed to consist of a substrate of pure Si, a SiO_2 layer (T_{SiO2}^{model}) and an organic contamination containing only the elements C and H (T_{org}^{model}). The principle of this method is presented only briefly; for details and other examples of application we refer to [77]. Within the model, simple exponential attenuation of the XPS signals is assumed. The intensity of the C1s signal is expressed in terms of the thickness T_{org}^{model}, the atomic density of the organic contamination, the XPS cross section or the sensitivity factor of the C1s line, the attenuation length L_C(C1s) for C1s electrons in the organic top layer and a number of instrumental parameters like the X-

ray flux, the transmission function, the detector efficiency for a given kinetic energy E_i and the correction factor for the asymmetry effect. Similar expressions can be derived for the intensity of the O1s signal, the intensity of the Si2p signal originating in the SiO_2 layer and the Si2p signal coming from the e-Si substrate. For the attenuation in the organic contamination we adopted values for the inelastic mean free path given by [78]. Elastic scattering in the SiO_2 layer and in the substrate was taken into account by using values for the attenuation length, taken from [68]. This provides altogether four equations with four unknown parameters: the thickness of the organic contamination T_{org}^{model}, the thickness of the SiO_2 layer T_{SiO2}^{model}, the concentration ratio c_O / c_{Si4+} in the SiO_2 layer and the X-ray flux. Reversal of these equations provides expressions for T_{org}^{model}, T_{SiO2}^{model} and c_O / c_{Si4+} in terms of the measured intensities. This model analysis has been applied to the present set of XPS analyses, all for Θ = 34° and azimuth = 22.5°. The sensitivity factor for the Si2p peak was chosen such that for sample L (not contaminated "infinitely thick" SiO_2) the ratio c_O / c_{Si4+} = 2.0 was obtained. The intensity of the Si2p peak of elementary signal depends - due to crystal effects - upon the measuring geometry (Θ and azimuth) and is also reduced by intrinsic plasmon losses [79]. To take these effects into account, the measured signal of the Si substrate was divided by $R_{0,th}/R_{0,exp}$ = 0.65 with $R_{0,th}$ = 0.529 (see [68]) and $R_{0,exp}$ = 0.81 (15). The resulted T_{SiO2}^{model}, T_{org}^{model} and the sum of these two thicknesses are summarized in Table 4.4.4. the concentration ratio C_O / C_{Si4+} in the SiO_2 layer is also reported.

Sample	Optical thickness	Standard equation (4)	Quases – Tougaard	Model calculation			Ratio
	T_{SiO2}^{opt}	T_{SiO2}^{std}	T_{SiO2}^{QT}	T_{SiO2}^{model}	T_{org}^{model}	$T_{SiO2}^{model} + T_{org}^{model}$	C_O / C_{Si4+}
A	0.14	0.00	0.06	0.04	0.28	0.32	-
B	0.50	0.29	0.73	0.36	0.16	0.53	3.0
C	1.00	0.72	1.14	0.78	0.15	0.93	2.6
D	1.01	0.77	1.35	0.83	0.18	1.01	2.5
E	1.07	0.56	1.17	0.66	0.19	0.85	2.8
F	1.41	1.21	1.93	1.24	0.14	1.38	2.4
G	1.42	1.19	1.84	1.21	0.18	1.39	2.3
H	2.00	1.86	2.76	1.85	0.14	1.99	2.2
I	2.20	1.94	3.10	1.92	0.09	2.02	2.2
J	2.51	2.38	3.39	2.35	0.14	2.49	2.2
K	3.20	3.09	4.71	3.03	0.11	3.14	2.1

Table 4.4.4: *Thickness of the SiO_2 layers (in nm) according to the analysis of the XPS spectra with the standard equation (4), with Quases-Tougaard [73] and with the model calculation. The measured optical thicknesses are also reported.*

The thicknesses of the ultra-thin SiO_2 layers obtained with these various techniques are summarized in Table 4.4.4 and compared to the measured optical thicknesses ($T_{SiO_2}^{opt}$).

A comparison between $T_{SiO_2}^{opt}$, $T_{SiO_2}^{std}$ and $T_{SiO_2}^{model} + T_{org}^{model}$ is presented in Figure 4.9(a). It is interesting to see, that the modeled thickness ($T_{SiO_2}^{model} + T_{org}^{model}$) is in very good agreement with $T_{SiO_2}^{opt}$. Apparently, the optical measurements determine the thickness of the combination of SiO_2 layer and the organic contamination. The thickness obtained with the standard equation (4) is in good agreement with the results of the model analysis: the difference is on average 0.02 nm, the largest difference being 0.1 nm.

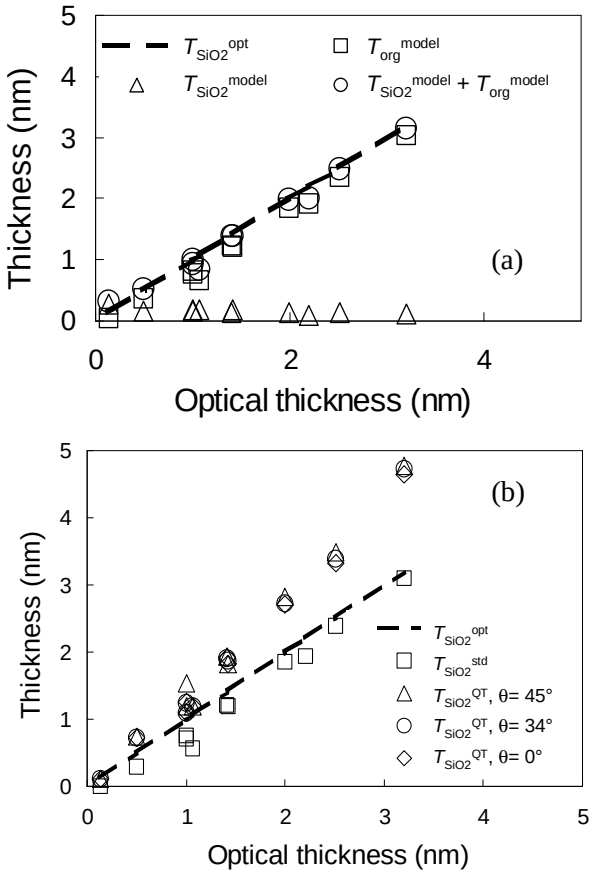


Figure 4.9: *Comparison of the SiO_2 layers thickness as determined from XPS results and from optical measurements. (a) Comparison between optical measurements (T_{SiO2}^{opt}) and thickness resulted from the model analysis (T_{SiO2}^{model}, T_{org}^{model}, $T_{SiO2}^{model} + T_{org}^{model}$). (b) Comparison between optical measurements and calculated thickness, T_{SiO2}^{std}, using the standard equation (14) and of the analysis of extended O1s peaks with Quases–Tougaard, T_{SiO2}^{QT}, [73] for • = 0°, 34° and 45°.*

The results of the Quases-Tougaard approach were in all cases above the results according to the standard equation, the model calculations and the optical thickness, the difference being on average 36 %. The results of the Quases-Tougaard analysis obtained at different angles are in very good mutual agreement, the differences being less than 0.2 nm (Figure 4.9(b)).
The concentration ratio O/Si^{4+} in the layers is within the experimental accuracy close to 2.0 for $T_{SiO2}^{model} > 2.5$ nm, but increases when the thickness decreases (Table 4.4.4 and in Figure 4.10). This can definitely not be attributed to an oxygen component in the organic contamination, as the C1s peak was, in all cases, corresponding to a aliphatic hydrocarbon without a detectable fraction of C-O bonds and because the thickness of the organic contamination was always less than 0.2 nm.

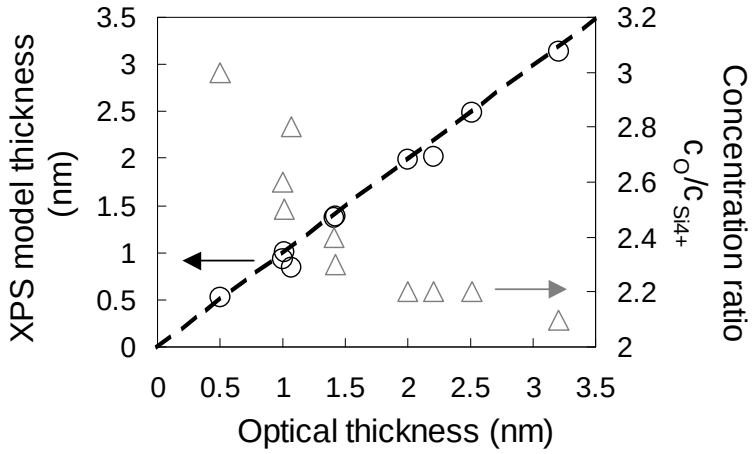


Figure 4.10: *Concentration ratio (C_O / C_{Si4+}) in the SiO_2 layer and total thickness from XPS model ($T_{SiO2}^{model} + T_{org}^{model}$) plotted as a function of the optical thickness.*

The thickness of the SiO_2 layer (T_{SiO2}^{model}) obtained from the XPS model have then been compared to thicknesses obtained with other techniques, namely RBS and TEM. The RBS results were, after a correction for the adhesion of oxygen during the RBS measurements, in good agreement with the results of the optical measurements and the model calculation, the difference being on average less than 0.1 nm (Table 4.4.5).

The thicknesses measured on TEM cross-section was determined from at least three HR images taken from different areas on the wafer (at least a few µm apart). The magnification was calibrated on every image studied; the calibration factor that was determined for each image appeared to be nearly constant for all images studied, illustrating the intrinsic reproducibility of the instrument. The thickness was determined using the technique described in section 4.2.5 of this chapter. Note that it was not possible to measure the thickness of ultra-thin samples A and B as well as E from TEM cross section images.

For layers with a thickness above 2.5 nm, the results are in agreement within the experimental accuracies. Yet, for layers with a thickness below 2.5 nm, the TEM data are significantly larger than the other values (optical, XPS, RBS), as presented in Table 4.4.5 and in Figure 4.11. The deviation might tentatively be explained by assuming that the density of these ultra thin SiO_2 layers is less than the density of bulk SiO_2. This will clearly influence the thickness as determined by RBS; also the attenuation lengths in XPS are expected to increase when the density of a material decreases. Yet, as can be seen in Table 4.4.5, the deviation for sample D is nearly 40 %, and it seems unlikely that the density of SiO_2 in these layers is 40 % less than the density of bulk SiO_2.

A different tentative explanation is as follows. According to [80] and [81], the electron energy loss spectroscopy (EELS) spectrum of the oxygen atoms at the SiO_2/Si interface is different from that of bulk SiO_2, indicating that the chemical environment of the oxygen atoms at the interface is different from bulk SiO_2. Further, the thickness of the SiO_2 layer that is obtained by measuring the O K edge as a function of position across the SiO_2 layer is significantly larger than the optical thickness (see Figure 3 in [80]: the optical thickness is 1 nm whereas the total O signal has a full width at half maximum, FWHM, of 1.6 nm). The results suggest that some of the surface oxygen atoms "penetrate" into the e-Si; this may give rise to a seemingly thicker SiO_2 layer in TEM images, as electron energy loss effects influence the intensity of the transmitted beam. For a sample in which the optical thickness of the SiO_2 is 1.8 nm, the FWHM of the total O signal is 2.1 nm. Apparently, the penetration of interfacial oxygen into the silicon decreases as a function of increasing thickness of the SiO_2 layer. This may explain the disappearance of the discrepancy between "TEM thickness" and "XPS or RBS thickness" as a function of increasing SiO_2 thickness. Yet, the reason why the "penetration" effect at the SiO_2/Si interface depends upon the thickness of the oxide layer for oxide thinner than 2.5 nm remains unknown.

Sample	**Optical**	**RBS**		**TEM**	
	T_{SiO2}^{opt}	**O atoms [10^{15} / cm^2]**	T_{SiO2}^{RBS}	T_{SiO2}^{TEM}	T_{SiO2}^{model}
A	0.14	1.12	0.03		0.04
B	0.50	2.4	0.3		0.36
D	1.01	5	0.86	1.37	0.83

E	1.07	4.7	0.81		0.66
F	1.41	7.62	1.43	1.68	1.24
G	2.00	10.1	1.96	2.12	1.85
J	2.51	12.5	2.48	2.43	2.35
K	3.20	16	3.23	3.33	3.03

Table 4.4.5: *Thickness of the SiO_2 layers (in nm) according to optical measurement (T_{SiO2}^{opt}), RBS (T_{SiO2}^{RBS}), TEM (T_{SiO2}^{TEM}) and the XPS model calculation (T_{SiO2}^{model}). Amount of oxygen at the surface according to the RBS measurements are also summarized for all studied samples.*

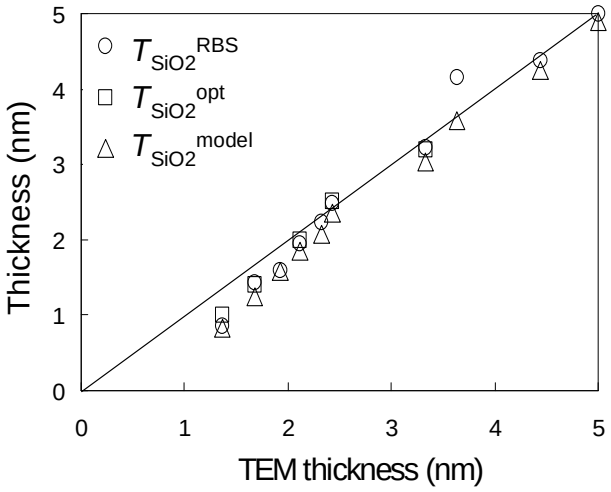


Figure 4.11: *Thickness of the SiO_2 layer according to RBS (T_{SiO2}^{RBS}), XPS (T_{SiO2}^{model}) and optical measurement (T_{SiO2}^{opt}), plotted as a function of the thickness according to TEM measurements (T_{SiO2}^{TEM}).*

4.3.3 Conclusions

The thickness of ultra-thin oxide layers ranging from 2 down to 0.14 nm has been determined using various techniques. We have shown, that good agreement between optical measurement, standard equation for XPS (4), model analysis of XPS ($T_{SiO2}^{model} + T_{org}^{model}$) and RBS. Model analysis of XPS results demonstrated in a quantitative way that deviations from the optical results are a consequence of surface contaminations of the samples with hydrocarbons. Yet, the thickness of this organic layer is very thin and therefore the ellipsometry technique seems to be a reliable measurement of the thickness of ultra-thin SiO_2 films. Optical measurement has the main advantage to be fast and cheap as compared to XPS or RBS techniques.

Quases-Tougaard analysis of extended O1s peaks has also been applied. The results were found to be independent of crystal effects, as expected. Yet, the thickness of the SiO_2 layers was approximately 36 % higher than the thickness determined with the other methods.
TEM results are in agreement for thickness exceeding 2.5 nm, but in samples with SiO_2 layers less than 2.5 nm the TEM thickness is larger than the optical, RBS or XPS thickness. Together with this effect we have observed that the concentration ratio O/Si^{4+} in these layers is larger than 2.0 and increases when the SiO_2 thickness decreases. These effects may be related to penetration of interfacial oxygen into silicon, as observed using EELS in ultra thin gate oxides of SiO_2.

4.4 Physical characterization of ultra-thin plasma nitrided oxides

In this section, the physical properties of ultra-thin plasma nitrided oxides formed with the DPN process (described in chapter 2) are characterized. There are several essential parameters to characterize when optimizing ultra-thin plasma nitrided oxides. The N concentration and distribution is of first importance as it is strongly linked to electrical parameters such as the reliability of the film and the mobility of carriers in the transistor channel [82], [83], [84]. An understanding of the chemical bonding configuration of N within the ultra-thin plasma nitrided film is therefore essential in the tuning process of the N profile within the layer. This has been studied with XPS analysis.
One of the greatest challenges involved in determining the scalability of any process beyond the sub 3.0 nm range is how to accurately measure the film thickness. For the DPN process, the physical thickness of the base oxide is measured optically, using single wavelength ellipsometry. We have shown previously that optical thickness of this SiO_2 layer is an accurate measurement. However, once nitrogen is incorporated into the film, the dielectric constant and, correspondingly, the index of refraction, have increased, making optical measurements of the nitrided film unreliable. The use of the base oxide thickness as a measure of the final dielectric thickness neglects any physical changes occurring during the DPN process, such as physical thickening, oxide re-growth or densification. TEM can be used to determine the thickness of any hybrid layer. However, as mentioned previously in this chapter, because of the difficulty to define accurately the interfaces of SiO_2 (or SiON) with the Si bulk and with the Al capping layer, the resolution accuracy of the TEM thickness is ± 0.1 nm, which is a few percent of the film thicknesses in the range of interest. Moreover, the very time consuming sample preparation required prior to TEM analysis makes this technique not adapted for the optimization of plasma nitrided films. Therefore alternate approaches to determine the film thickness are here investigated using XPS, TOFSIMS and RBS.

The incorporation of nitrogen into the SiO_2 film results in an increase in the film density because silicon nitride (Si_3N_4) is more dense than silicon dioxide (SiO_2). In the case of TOFSIMS measurement, a higher density film results in a slower sputter rate, thus making the films look artificially thicker. The TOFSIMS measurements obtained on our heavily nitrided samples were thus inaccurate. It was therefore of first importance to calibrate our TOFSIMS measurements. A concentration and depth scale calibration method has been developed for ultra-thin SiO_xN_y layers on Si. The method has been tested over a wide range of compositions by comparison of integrated N and O dose with RBS measurements [66].

4.4.1 N incorporation in ultra-thin plasma nitrided oxide films

The possible N bonding configurations for N incorporation in the SiO_2 matrix are given in Figure 4.12. As presented in section 4.2.2 of this chapter, XPS analysis is used to determine the chemical bonding configuration of atoms in measuring the binding energies of the elements present in the studied sample. For the study of plasma nitrided oxides, the binding energies of N, O and Si as SiON is of interest. The N bond structure is identified with peak assignments for N1s in the bond configurations of interest given in Table 4.4.6 [82], [87]. Calculations have shown that the N1s binding energy shifts by ~1.8 to 2.0 eV for each additional O atom which is bonded to nitrogen.

Bond Configuration	N Peak Position (eV)
Si≡N	397.3 - 398.5
O-Si=N	399.4 – 400.4
Si=N-O	401.8 – 402.6

Table 4.4.6: *XPS Peak Assignment for N in oxynitrides for possible bonding configurations [86] and [87].*

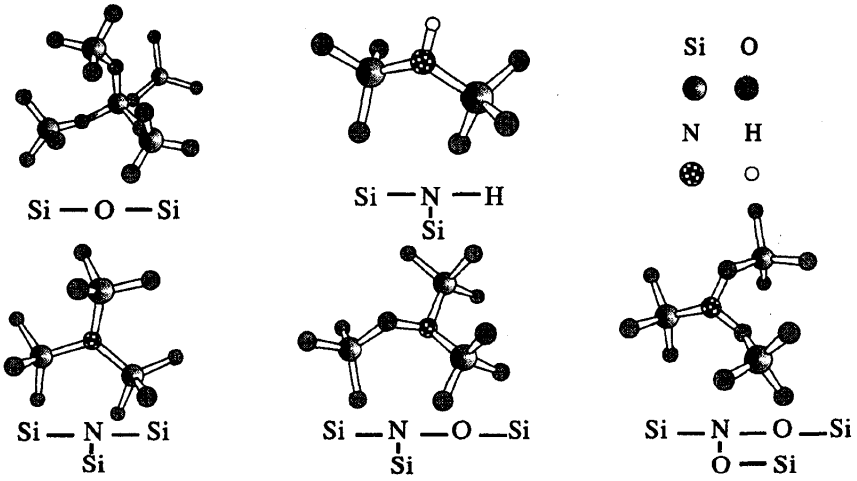


Figure 4.12: *Cluster models for the representative bond created with the incorporation of nitrogen atom into the silicon dioxide film. Terminating hydrogen atoms are omitted for clarity. Note that N bonded only to Si is most stable configuration for N incorporation.*

The chemical bond configuration of N in the SiON matrix has been here studied after the full DPN process step [88]. Photoelectron spectra of plasma nitrided oxides on silicon were measured using the system described in section 2.2.2. Only one N1s peak was observed, as presented in Figure 4.13(a). This result yields one type of N bonding configuration in plasma nitrided oxides. The binding energy measured for the N1s peak is 397.6 ± 0.1 eV. This indicates that N atoms are mainly bonded to Si atoms creating Si≡N bonds (Table 4.4.6). The chemical bond configuration of N is therefore very close to that of N in Si_3N_4 [89]. The measured Si spectrum exhibits two peaks (see section 4.2.2), as shown in Figure 4.13(b). The binding energy measured for the peak of the Si chemically bound to O and N (Si as SiON peak) is around 102 eV. In SiO_2, this peak is at 103.5 eV, in Si_3N_4 it is at approximately 101.8 eV. The binding energy measured is 532.8 eV which is very close to the binding energy of silicon dioxide (533.2 eV according to [86]). These results confirm that N atoms are mainly bonded to Si atoms and for Si≡N bonds.

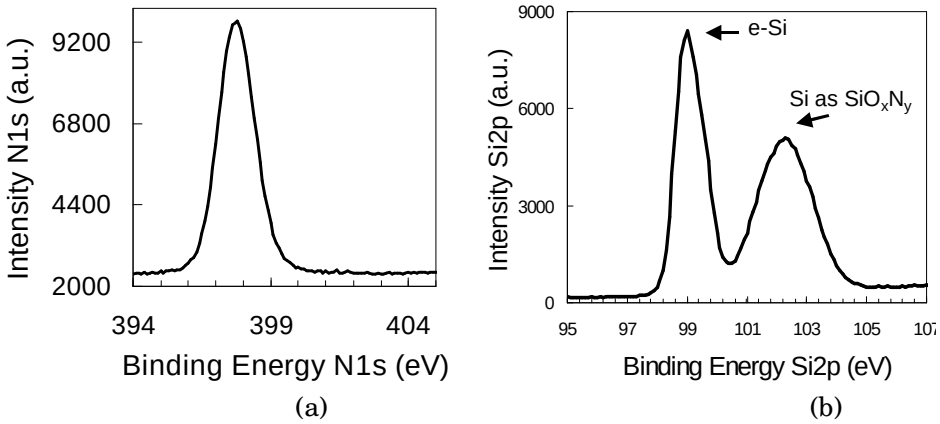


Figure 4.13: *(a) Scheme showing various possible O and N diffusion mechanisms that could occur in the SiON layer. (b) Spectra of the N1s element after full DPN process. Only one peak is measured at a binding energy close to N as Si_3N_4, indicating that most of the N atoms are bonded to Si atoms.*

Moreover, the position of the N1s and Si as SiON peaks has been studied for plasma nitrided oxides with various nitrogen content. The N1s and Si2p as SiON peak positions are plotted as a function of the ratio of N over O in Figure 4.14(a) and (b), respectively. It can be observed that both the N1s and Si2p as SiON peaks shift to lower binding energies when increasing the effective amount of N. this indicates that more N atoms are incorporated in the films and are bonded to Si atoms to form Si≡N bonds.

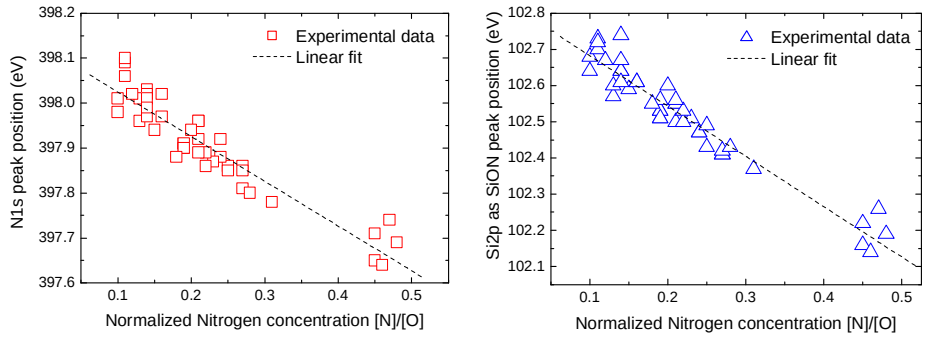


Figure 4.14: *Binding energies of the N1s and Si2p as SiON peaks as a function of the normalized nitrogen concentration. The data were obtained from several series of ultra-thin SiON layers, with different thicknesses; also the DPN times were varied as well as the atmosphere of the ambient during nitridation.*

4.4.2 N concentration in ultra-thin plasma nitrided oxide films

The amount of N in ultra-thin plasma nitrided oxides has been measured with different techniques, namely TOFSIMS, RBS and XPS. For TOFSIMS analysis, a calibration of the profile has been developed and the resulted concentrations have been compared with the ones obtained with other techniques.

4.4.2.1 Concentration scale calibration for TOFSIMS measurement

The sum of all Si_2N^+ intensities, I_{Si2N}, is defined as:

$$I_{Si2N} = \frac{I_{70}}{0.851} \tag{4.7}$$

where I_{70} is the intensity measured in an interval around the mass of $^{28}Si_2N$ (0.851 because 85.1% of all Si_2N clusters have mass 70). Around mass 72, $^{28}Si^{30}SiN^+$ and $^{29}Si_2N^+$ interfere with $^{28}Si_2O^+$. Therefore, the Si_2O^+ intensity (5.94% of all Si_2N-clusters have mass 72) is defined as:

$$I_{Si2O} = \frac{I_{72} - I_{70} \times (0.0594/0.851)}{0.851} \tag{4.8}$$

With Ar primary ions, there are no other significant mass interferences. The selected reference signal is:

$$I_{Si2} = \frac{I_{56}}{0.851} \tag{4.9}$$

The intensity I_{Si2N} (and equivalently I_{Si2O}) can be expressed as:

$$I_{Si2N} = x_{Si}^2 x_N Y P_{Si2N} \Phi_p \tag{4.10}$$

where x_{Si} and x_N are the silicon and nitrogen relative concentrations (for all samples at all depths, we assume $x_{Si} + x_N + x_O = 1$), Y is the total sputter yield, P_{Si2N} is the chance that two Si atoms and one N atom form a Si_2N^+ ion that is transmitted and detected by the instrument, and Φ_p is the primary ion flux.

A suitable reference signal in this case is the intensity I_{Si2} of Si_2^+ (4.10) because the ratio I_{Si2N}/I_{Si2} can be expected to increase monotonously with x_N, which is not necessarily the case for I_{Si2N} as such (since x_{Si} decreases as x_N increases).

$$I_{Si2} = x_{Si}^2 Y P_{Si2} \Phi_p \tag{4.11}$$

The ratio I_{Si2N}/I_{Si2} and I_{Si2N}/I_{Si2} are further called respectively I_N and I_O. The relative sensitivity factors, S_N is defined according to:

$$S_N = \frac{x_N}{I_N} = \frac{P_{Si2}}{P_{Si2N}} \tag{4.12}$$

S_O is defined similarly.

We have determined S_N and S_O in Si and Si_3N_4 from well-calibrated implants: $S_{N\ in\ Si}$= 0.0167(4.8), $S_{O\ in\ Si}$=0.155(4.15), $S_{O\ in\ S3N4i}$=0.098(4.15) and the stoichiometric composition of Si_3N_4 (x_N=4/7): $S_{N\ in\ Si3N4}$= 0.041(4.2).

For SiO_2, we have done the same, but we found non-linear behavior of I_O and I_N for SiO_2 with a few percentage of N. Since we are particularly interested in quantification of higher N contents, we have used here values that describe SiO_2 with ~5% N: $S_{N\ in\ SiO2}$= 0.020(4.3) and $S_{O\ in\ SiO2}$=0.20(4.3). As a consequence, dilute N in SiO_2 is significantly underestimated in relative sense, but in absolute sense the errors are very small.

To calibrate intensities from an unknown composition in between the basic materials, the composition is considered a mixture of Si, SiO_2 and Si_3N_4:

$$Si_u(SiO_2)_v(Si_3N_4)_w \tag{4.13}$$

where $u+v+w=1$ (note that $x_o=2/3v$ and $x_N=4/7w$), and S_N and S_O are linearly interpolated: $S_N= uS_{N\ in\ Si} + vS_{N\ in\ SiO2} + wS_{N\ in\ Si3N4}$ and $S_O= uS_{O\ in\ Si} + vS_{O\ in\ SiO2} + wS_{O\ in\ Si3N4}$. Then explicit expressions for x_N and x_O (or for u, v and w) in terms of I_N and I_O can be derived:

$$x_N = I_N \frac{8S_N^{Si} + 12I_O(dS_O^{Si} - cS_N^{Si})}{8 - 12cI_O - 14bI_N + 21(bc - ad)I_O I_N} \tag{4.14}$$

$$x_O = I_O \frac{8S_O^{Si} + 14I_N(aS_N^{Si} - bS_O^{Si})}{8 - 12cI_O - 14bI_N + 21(bc - ad)I_O I_N} \tag{4.15}$$

where $a=S_{O\ in\ Si3N4}-S_{O\ in\ Si}$, $b=S_{N\ in\ Si3N4}-S_{N\ in\ Si}$, $c=S_{O\ in\ SiO2}-S_{O\ in\ Si}$ and $d=S_{N\ in\ SiO2}-S_{N\ in\ Si}$. Of course, such a linear variation of sensitivity with x_N and x_O is an approximation, but since the variation of the sensitivity factors over the full range of materials is small (maximum a factor 2.5, which is much less than e.g. for CsN^+ and CsO^+ clusters [65]), the choice is not so important.

For a concentration calibration in terms of atomic density, the concentrations x_N and x_O must be multiplied by total atomic density of the mixture N, which can also be linearly interpolated between the basic materials: $N = uN^{Si}+vN^{SiO2}+wN^{Si3N4}$.

4.4.2.2 Benchmark of various techniques to measure the N concentration

TOFSIMS analysis using the above calibration method has been tested over a wide range of samples with various compositions by comparison of integrated N and O dose with RBS measurements. As sown in Figure 4.15, the ratios of the nitrogen and oxygen doses calculated with SIMS and RBS yield very good agreement.

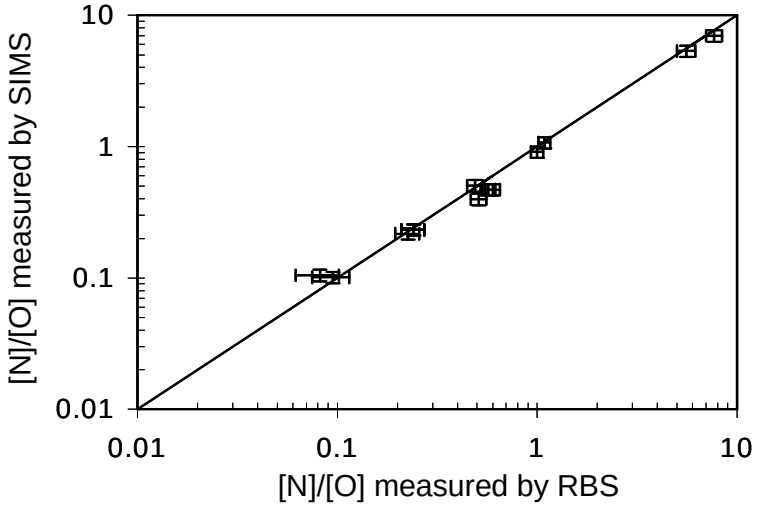


Figure 4.15: *Comparison of the ratio of the nitrogen and oxygen integrated dose as determined by RBS and SIMS for a wide range of layer compositions (line indicates a 1/1 agreement).*

As described in section 2.2 of this chapter, XPS can be also used to measure the concentration of species. A comparison of the concentration of N obtained with XPS and with RBS has been made over a wide range of ultra-thin plasma nitrided thin films (Figure 4.16). Again a good agreement between the two techniques can be observed for a wide range of N concentration.

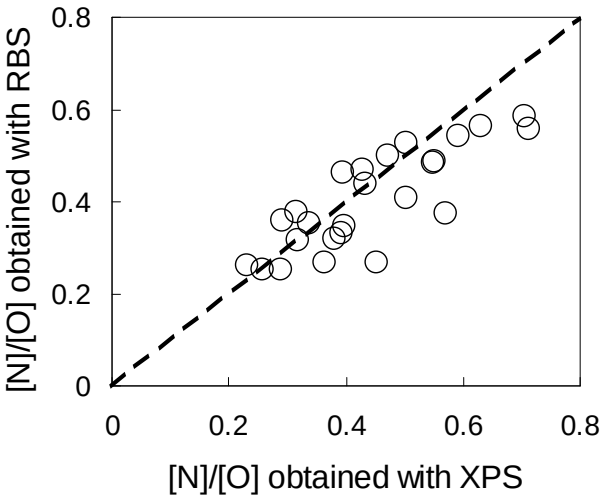


Figure 4.16: *Comparison of the ratio of the nitrogen and oxygen integrated dose as*

determined by XPS and RBS for various ultra-thin plasma nitrided oxides.

Finally, the D2R method, as described in section 2.1, has been applied on a wide range of ultra-thin plasma nitrided oxide films having a physical thickness below 2 nm. The resulted data has been compared to the nitrogen content within the film as measured from XPS analysis. Although the D2R technique is a method that evaluates the interfacial N concentration (see section 4.2.1), a very good agreement was found between the D2R parameter and the total N content in the film, as shown in Figure 4.17.

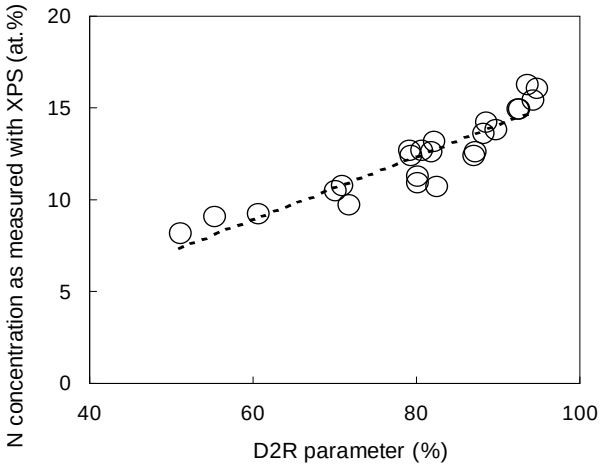


Figure 4.17: *Comparison of the N concentration measured with XPS and calculated using the D2R method (described in section 2.1 of this chapter) for ultra-thin oxides (physical thickness is below 2 nm).*

From these results, it can be concluded that TOFSIMS (using the developed concentration calibration), RBS and XPS can be used to measure accurately the amount of nitrogen in thin SiON layers and this for a large range of N concentration. A linear correlation between the N concentration as measured with XPS and the D2R parameter has been obtained for N concentration ranging from 8 up to 16 atomic percent.

4.4.3 N distribution profile within ultra-thin plasma nitrided oxides

Measuring the N profile within an ultra-thin nitrided oxide film is a real challenge. Indeed, for most of the techniques, the depth resolution is above 1 nm which is more than 50 % of the total physical thickness of the film. TOFSIMS analysis have been performed on ultra-thin plasma nitrided oxide and the N, O and Si profiles have been investigated, as shown in Figure 4.18. The SiON/Si interface was determined from XPS measurements (see next section of this chapter).

It can be observed that the peak of the N profile is not located at the interface with the Si substrate, which is in accordance with published results on the N profile of plasma nitrided films [90], [91], [92]. However, most of these papers are presenting results showing that the N peak concentration is located at the top surface. These studies were performed for rather thick plasma nitrided oxides (physical thickness > 2nm) and their conclusions on the location of the N peak at the top surface cannot be drawn in our case. Indeed, in our study of ultra-thin plasma nitrided (physical thickness < 2nm), the depth resolution limit of our TOFSIMS measurements is about 50 % of the total physical thickness of the SiON layer. Moreover, surface contamination and initial sputtering effects make the quantitative measurement of the peak position questionable. However, the good correlation observed between the total nitrogen content with the D2R parameter (see Figure 4.17) might indicate that the N atoms are not located at the top surface but are more homogeneously distributed over the film.

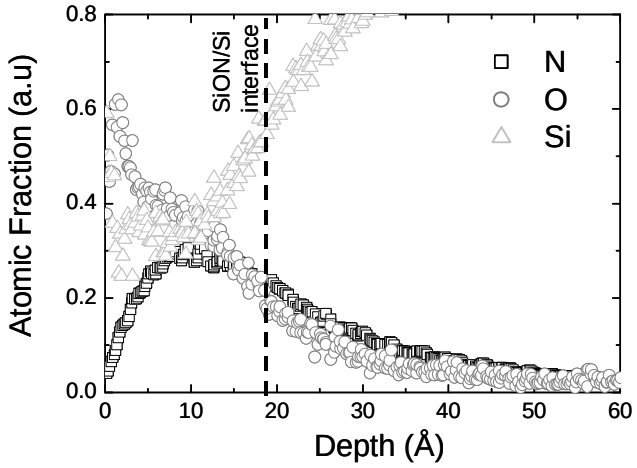


Figure 4.18: *N, O and Si profiles as measured with TOFSIMS of an ultra-thin plasma nitrided film. The physical thickness of the film is about 1.8 nm.*

4.4.4 Thickness measurement of ultra-thin plasma nitrided oxide films

XPS analysis was used to evaluate the thickness of the SiON layers. The standard approach to determine the thickness of the SiO_2 layer is to use the standard equation (4.4) as detailed in section 3. However, this equation is taking the attenuation length of a pure SiO_2 film (L_{SiO2} in (4.4)). The attenuation length (L) of the SiON film was investigated. In the mixtures Si-O-N, values for L[O1s], L[N1s] and L[Si2p] can be obtained using the expression:

$$L = x^{*}L(\text{pure } SiO_2) + (1\text{-}x)\, L(\text{pure } Si_3N_4) \qquad (16)$$

The calculated L in SiON versus the ratio [N]/[O] is shown in Figure 4.19. It can be seen that between pure oxide and nitride film, the attenuation length (L[Si2p]) does not vary dramatically: from 34.48 Å for SiO_2 to 29.66 Å for Si_3N_4. Therefore, a small change in the N content of the SiON layer will not induce a dramatic error (< 10 %) in the estimation of the thickness of the dielectric if the attenuation length is taken as pure oxide (L[Si2p]=L[SiO2]=34.48 Å).

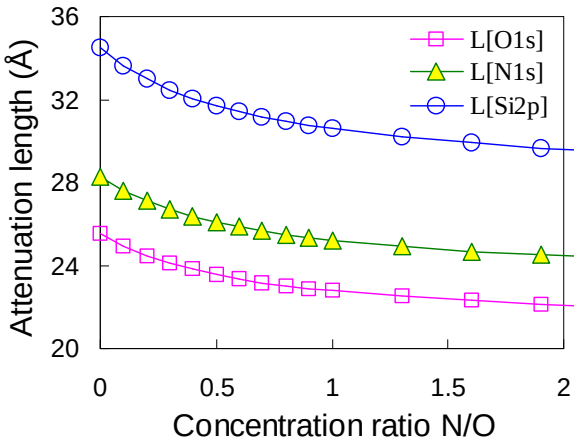


Figure 4.19: *Calculation of the attenuation lengths L[Si2p], L[O1s], L[N1s] versus the ratio [N]/[O].*

The thickness of SiON films was also estimated using RBS or TOFSIMS analysis, as already presented in this chapter. For TOFSIMS analysis, a depth calibration has been developed.

4.4.4.1 Depth scale calibration for TOFSIMS measurement

We have used optical profilometry of the final craters for depth scale calibration (stylus profilometry is also possible, but it is difficult to locate the craters and to ensure a trace exactly through the crater centres). Optical profilometry with a true phase-shifting algorithm of an ultra-thin oxide or nitride layer on Si measures (a little bit more than) the depth sputtered into the Si substrate (Figure 4.20) [63]. Thus, the erosion rate of pure Si, z^{Si}, can be determined from the optical depth d_{opt} and the period of time sputtered after passing through the interface.
For the interface, we take the sputter time t where the fraction of Si (u) is equal to 50 % of the sum of the O and N fractions:

$$u(t) = v(t) + w(t) = \tfrac{1}{2} \quad (x_{Si} \sim 0.7) \qquad (4.17)$$

To account for changes in erosion rate, the ratios $r^{Si3N4}=z^{Si3N4}/z^{Si}=0.85(4)$ in Si_3N_4 and $r^{SiO2}=z^{SiO2}/z^{Si}=0.93(4)$ in SiO_2 with 500 eV Ar^+ at 45º have been determined from stylus profilometry of deep craters in the pure materials. z as a function of t is taken as:

$$z(t) = (u(t) + v(t)r^{SiO_2} + w(t)r^{Si_3N_4}) \times z^{Si} \tag{4.18}$$

Unlike Cs or O, little Ar is incorporated in the target, making such linear composition-dependence plausible.
Finally, the depth $d(t)$ at the sputter times t_i is expressed as:

$$d(t_i) = d(t_{i-1}) + z(t_i)(t_i - t_{i-1}) \tag{4.19}$$

starting with $d(t_0) = 0$ at $t_0 = 0$.

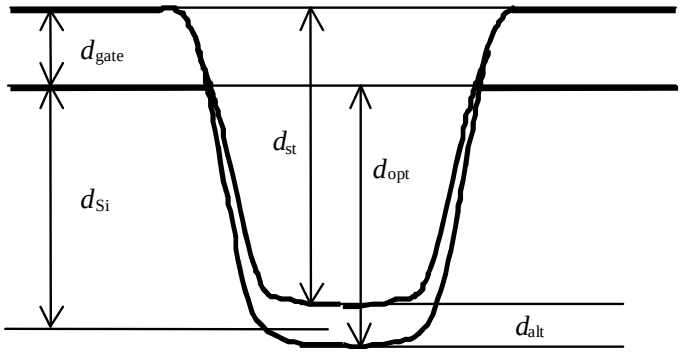


Figure 4.20: *Explanation of crater depths measured using stylus profilometry (d_{st}) or optical profilometry (d_{opt}) in case of a gate dielectric (d_{gate} ~ few nm's) on Si. The amount of Si sputtered is equivalent to a thickness d_{Si} and after exposure to air the crater bottom is oxidized to an altered layer with d_{alt} ~ 1 nm. For simplicity, d_{opt} is drawn from interface to interface although this is not exactly true.*

With this procedure, the depth scale accuracy improves by prolonging the profile, because the crater depth measurement improves and because the depth scale becomes less sensitive to the choice of the interface. With sufficiently deep profiles, this method is more accurate than using the layer thickness measured by ellipsometry for example or by using a depth scale standard. A typical depth profile of a gate dielectric, thus obtained, is given in Figure 4.21.

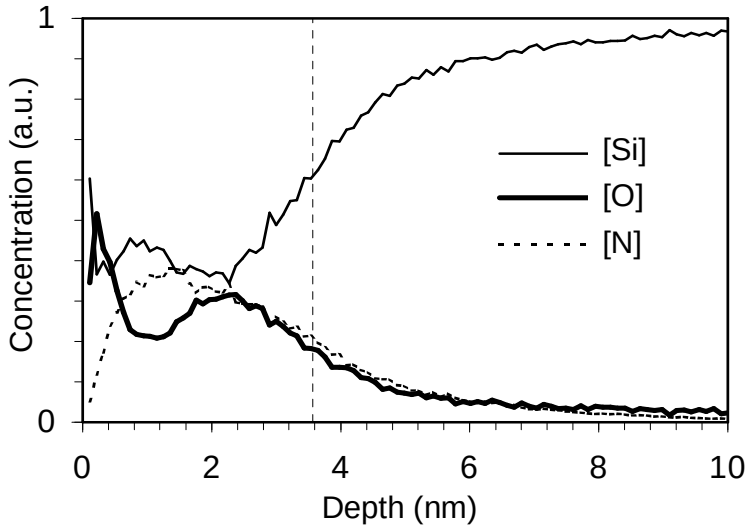


Figure 4.21: *Si, O and N concentration depth profiles of an LPCVD nitride layer on Si. The vertical dashed line (d = 3.57 nm) indicates the depth where a hypothetical split into fractions u of Si, v of SiO2 and w of Si3N4 reaches u = v+w = ½. The total layer contains about equal amounts of O and N, with O having some preference for the surface and the interface.*

4.4.4.2 Benchmark of various techniques to measure the physical thickness of ultra-thin plasma nitrided oxides

A benchmark of all of these techniques has been made for various SiON gate dielectrics ranging from 2.2 down to 1.4 nm (Figure 4.22). Note that TOFSIMS profiles were measured only on a limited amount of samples. It can be observed that for SiON films thicker than 1.5 nm, XPS, RBS and TOFSIMS analysis are in good agreement. However, for films thinner than 1.5 nm, values obtained with XPS are sensibly smaller than RBS or TOFSIMS ones. While RBS and TOFSIMS are accurate techniques to measure the film composition and stochiometry, their poor depth resolution provide reliable thickness measurement in the sub-2 nm regime (see Table 4.4.1 and [82]).

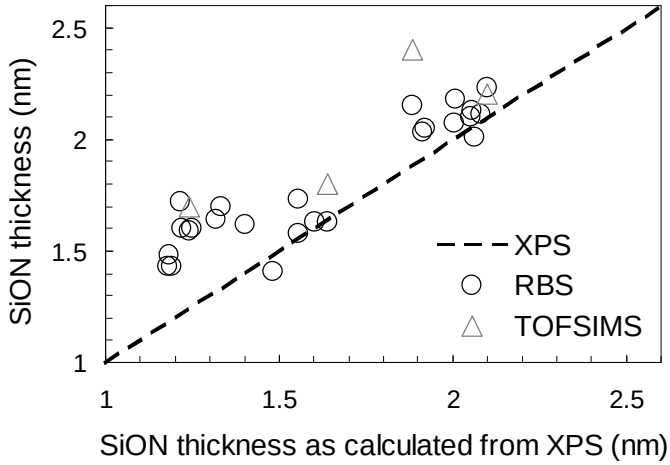


Figure 4.22: *Comparison of the thickness of various SiON thin films obtained with XPS, and TOFSIMS techniques.*

The physical thickness of ultra-thin plasma nitrided films can be also calculated from ellipsometry measurement using the D2R technique. It was shown in Figure 4.17 that the N concentration measured with XPS and the D2R parameter were in good agreement for N concentration in the range of 8 to 16 atomic percent. In this particular range, the physical thickness based on D2R data was calculated and compared to the physical thickness calculated from XPS data. As expected, a good correlation between these two calculated thicknesses was obtained, as depicted in Figure 4.23.

This result shows that the physical thickness of ultra-thin plasma nitrided oxide films can be estimated from ellipsometry measurements using the D2R method and from the correlation with XPS data as shown in Figure 4.23, providing that the N content is within a certain concentration range, namely 8 to 16 atomic percent.

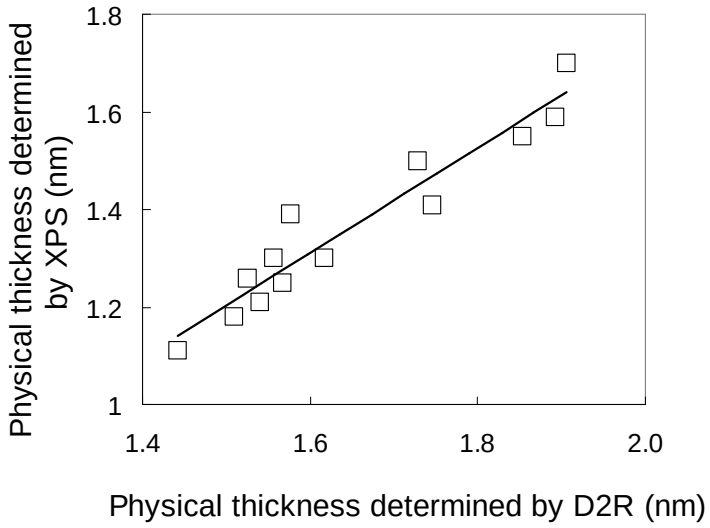


Figure 4.23: *Comparison of the physical thickness determined with XPS measurements and with the D2R method.*

4.4.5 Conclusions

The materials characterization of DPN nitrided dielectric films has been conducted using ellipsometry, TOFSIMS, TEM, RBS and XPS analysis. These techniques were chosen based on their availability and expertise.

It was found that while RBS and TOFSIMS are accurate techniques to measure quantitatively the composition and stochiometry of ultra-thin plasma nitrided films, they are not suitable for the measurement of the thickness of such layers. Cross-section TEM can be used for the determination of the physical thickness although its accuracy depends on the sample preparation and on the accuracy of the localization of the interface dielectric/Si substrate. XPS, however, has the advantage to provide not only a good measurement of the film composition but also of the chemical bonding configuration and the physical thickness.

For a quick estimation of the N content and physical thickness of ultra-thin plasma nitrided oxides, ellipsometry measurement in combination with the D2R method can be used. It was shown that the resulted N content and physical thickness was in good agreement with XPS analyses. This technique has also the major advantage to be cheap as compared to any physical analysis.

The nitrogen distribution and chemical bonding configuration of ultra-thin plasma nitrided films have been investigated using XPS. It was found that after the full DPN process, the N atoms are bonded mainly to Si atoms. N atoms seem to exchange exclusively with O atoms in the SiO_2 matrix and form Si≡N type of bonds.

4.5 References

[46] Charles Evans & Associates, analytical characterization service provider: www.cea.com.

[47] H. G. Tompkins and W.A.McGahan, *Spectroscopic Ellipsometry and Reflectometry*, John Wiley & Sons, New York, 1999.

[48] R.M.A. Azzam and N.M.Bashara, *Ellipsometry and Polarized Light*, North Holland Press, Amsterdam 1977, Second edition, 1987.

[49] R.W.Collins, D.E.Aspnes, and E.A. Irene, *Spectroscopic Ellipsometry*, Editors, Elsevier Science S.A., 1998, Lausanne, Switzerland; also appears as Vol. 313-314 Thin Solid Films, Numbers 1-2, 1998.

[50] D. Briggs and M. P. Seah, *Practical Surface Analysis*, John Wiley & Sons, New York, 1990.

[51] Physical Electronics USA, PHI 2000 MultiTechnique XPS system: www.phi.com.

[52] M. P. Seah and S.J. Spencer, *Ultrathin SiO2 on Si II. Issues in quantification of the oxide thickness*, Surface and Interface Analysis, Vol. 33, pp. 640-652, 2002.

[53] E. Bøgh and E. Uggerhoj, *Experimental investigation of orientation dependence of Rutherford scattering yield in single crystals*, Nucl. Instr. and Meth., Vol. 38, pp. 216-220, 1965.

[54] W. K. Chu, J. W. Mayer and M. A. Nicolet, *Backscattering Spectrometry*, Academic Press Inc., 1978.

[55] J.R. Tesmer and M. Nastasi, *Handbook of Modern Ion Beam Materials Analysis*, Materials Research Society, Pittsburg, 1995.

[56] RUMP - RBS Analysis and Simulation Package [v. 4.00(beta)], (c) 1988-1997 Michael Thompson, Larry Doolittle, (c) 1988-1997 Computer Graphic Service, Ltd. All rights reserved, Revision Level:

Version 0.950; L. R. Doolittle, Nucl. Instr. and Meth., B 9, p. 5008, 1985.

[57] H.A. Storms, K.F. Brown, and J.D. Stein, ????, Analytical Chemistry, Vol. 49, pp. 2023-2026, 1977.

[58] Y. Gao, *A new secondary ion mass spectrometry technique for III-V semiconductor compounds using the molecular ions CsM* $^{+}$, Journal of Applied Physics, Vol. 64, pp. 3760-3762, 1988.

[59] T. Hoshi, L. Zhanping, M. Tozu, and R. Oiwa, ????, Secondary Ion Mass Spectrometry SIMS XI, Chichester, UK John Wiley & Sons, pp. 269-275, 1998.

[60] J. Vlekken, T.D. Wu, M. D'Olieslaeger, G. Knuyt, W. Vandervorst, and L.D. Schepper, *Monte Carlo simulation of the formation of M2 - Molecular Ions sputtered from metallic materials*, Secondary Ion Mass Spectrometry SIMS XI, Chichester, UK John Wiley & Sons, p. 895, 1998.

[61] J. Vlekken, T.D. Wu, M. D'Olieslaeger, G. Knuyt, L.D. Schepper, and L.M. Stals, *Monte Carlo simulation of the formation of MCs/sup+/molecular ions*, International Journal on Mass Spectrometry Ion Processes, Vol. 156, pp. 61-66, 1996.

[62] K. Snowdon, ????, Nuclear Instrumentation Methods B, Vol. 9, pp. 132-134, 1985.

[63] S.A. Schwarz and C.R. Helms, *New models of sputtering and ion knock-on mixing*, Secondary Ion Mass Spectrometry SIMS-II. Proceedings of the Second International Conference, pp. 15-17, 1980.

[64] C.J. Han, M.M. Moslehi, C.R. Helms, and K.C. Saraswat, *Characterization of thermally nitrided SiO2 using Auger sputter profiling.* Journal of Vacuum Science and Technology A, Vol. 3, pp. 804-805, 1985.

[65] M.A. Douglas, S. Hattangady, and K. Eason, *Depth Profile Analysis of Ultrathin Silicon Oxynitride Films by TOFSIMS*, Journal of the Electrochemical Society, Vol. 147(5), pp. 1893-1895, 2000.

[66] J.G.M. van Berkum, M.J.P. Hopstaken, J.H.M. Snijders, Y. Tamminga, F.N. Cubaynes, *Quantitative depth profiling of SiOxNy layers on Si*, Applied Surface Science, 2003, pp. 414-417.

[67] M .A. Verheijen and J.G.M. van Berkum, *TEM analysis report 2003.0100 TE030001*, CFT internal report, Philips, 2003.

[68] M.P. Seah and S.J. Spencer, *Ultra-thin SiO2 on Si: II, Issues in Quantification of the Oxide Thickness*, Surface and Interface Analysis, Vol. 33, pp. 640-652, 2002.

[69] M.P. Seah and S.J. Spencer, *Ultrathin SiO2 on Si. I. Quantifying and removing carbonaceous contamination*, Journal of Vacuum Science and Technology A, Vol. 21, pp. 345-352, 2003.

[70] B. S. Semak, C. van der Marel and S. Tougaard, *Comparison of the Tougaard, ARXPS, RBS and ellipsometry methods to determine the thickness of thin SiO_2 layers*, Surface and Interface Analysis, Vol. 33, pp. 238-244, 2002.

[71] M. Seah, *Intercomparison of Silicon Dioxide Thickness Measurement Made by Multiple Techniques*, J. Vac. Sc. And Tech. A, in press, 2004.

[72] C. van der Marel, M.A. Verheijen and Y. Tamminga, R. H. W. Pijnenburg, N. Tombros, F. Cubaynes, *The thickness and composition of ultra-thin SiO_2 layers on Si*, J. Vac. Sc. Tech. A, in press, 2004.

[73] S. Tougaard, *Accuracy of the non-destructive surface nanostructure quantification technique based on analysis of the XPS or AES peak shape*, Surface and Interface Analysis, Vol. 26, pp. 249-269, 1998; and S. Tougaard QUASES-Tougaard: Software package for Quantitative Analysis of Surfaces by Electron Spectroscopy, version 4.4, 2000.

[74] Software package for the analysis of XPS results, CasaXPS, version 2.2.32.

[75] D.A. Shirley, *High-Resolution X-Ray Photoemission Spectrum of the Valence Bands of Gold*, Phys. Rev. B, Vol. 5, pp. 4709-4714, 1972.

[76] Z.H. Lu, M.J. Graham, D.T. Jiang and K.H. Tan, *SiO_2/Si(100) interface studied by Al K x-ray and synchrotron radiation photoelectron spectroscopy*, Appl. Phys. Lett., Vol. 63, pp. 2941-2943, 1993.

[77] C. van der Marel, *A multilayer approach for the quantitative analysis of XPS-results*, to be published in Surface and Interface Analysis.

[78] P.J. Cumpson, *Estimation of inelastic mean free paths for polymers and other organic materials: use of quantitative structure-property relationships*, Surface and Interface Analysis, Vol. 31, pp. 23-34, 2001.

[79] T. Katayama, H. Yamamoto, M. Ikeno, Y. Mashiko, S. Kawazu and M. Umeno, *Accurate Thickness Determination of Both Thin SiO_2 on Si and Thin Si on SiO_2 by Angle-Resolved X-Ray Photoelectron Spectroscopy*, Jpn. J. Appl. Phys., Vol. 38, pp. 4172-4179, 1999.

[80] D.A. Muller, T. Sorsch. S. Moccio, F.H. Baumann, K. Evans–Lutterodt and G. Timp, *The electronic structure at the atomic scale of ultrathin gate oxides*, Nature, Vol. 399, pp. 758-760, 1999.

[81] D.A. Muller and J.B. Neaton, *Structure and energetics of the interface between Si and amorphous SiO_2* in *Fundamental Aspects of Silicon Oxidation*, edited by Y.J. Chabal, Springer Verlag Berlin Heidelberg New York, pp. 219-246, 2001.

[82] M.L. Green, E.P. Gusev, R. Degraeve and E.L. Garfunkel, *Ultrathin ([less-than] 4 nm) SiO[sub 2] and Si--O--N gate dielectric layers for silicon microelectronics: Understanding the processing, structure, and physical and electrical limits*, Journal of Applied Physics, Vol 90, pp. 2057-2121, 2001.

[83] D. Bouvet, P.A. Clivaz, M. Dutoit, C. Coluzza, J. Almeida, G. Margaritondo and F. Pio, *Influence of nitrogen profile on electrical characteristics of furnace or rapid thermally nitrided silicon dioxide films*, Journal of Applied Physics, Vol 79, pp. 7114-7122, 1996.

[84] S.V. Hattangady, H. Niimi and G. Lucovsky, *Controlled nitrogen incorporation at the gate oxide surface*, Applied Physics Letters, Vol. 66, pp. 3495-3497, 1995.

[85] S.W. Novak, J.R. Shallenberger, D.A. Cole and J.W. Marino, *Structure and bonding in nitrided oxide films by SIMS and XPS. in Ultrathin SiO2 and High-K Materials for ULSI Gate Dielectric*, Symposium. Materials Research Society, San Francisco, CA, USA, pp. 579-586, 1999.

[86] J.F. Moulder, W.F. Stickle, P.E.Sobol, and K.D.Bomben, *Handbook of X-ray Photoelectron Spectroscopy*, Minneapolis, 1992.

[87] R.I. Hegde, B. Maiti, and P.J. Tobin, *Growth and film characteristics of N2O and NO oxynitride gate and tunnel dielectric.* Journal of the Electrochemical Society, Vol. 144(3), pp. 1081-1086, 1997.

[88] F. Cubaynes, J. Schmitz, C. van der Marel, H. Snijders, A. Veloso, A. Rothschild, *Plasma nitridation optimisation for sub 15Å gate dielectrics*, Proc. ECS Paris, PV 2003-02, pp. 595-604 , 2003.

[89] D. Bouvet, P.A. Clivaz, M. Dutoit, C. Coluzza, J. Almeda, G. Margaritondo, and F. Pio, *Influence of nitrogen profile on electrical characteristics of furnace- or rapid thermally nitrided silicon dioxide films*, Journal of Applied Physics, Vol. 79, pp. 7114-7122 ,1996.

[90] H. Niimi and G. Lucovsky, *Monolayer-level controlled incorporation of nitrogen in ultrathin gate dielectrics using remote plasma processing: Formation of stacked '' N– O– N'' gate dielectrics*, Journal of Vacuum Science and Technology B, Vol 17, pp. 2610-2621, 1999.

[91] R. Kraft, T. P. Schneider, W. W. Dostalik and S. Hattangady, *Surface nitridation of silicon dioxide with a high density nitrogen plasma*, Journal of Vacuum Science and Technology B, Vol 15, pp. 967-970, 1997.

[92] S. V. Hattangady, H. Niimi and G. Lucovsky, *Controlled nitrogen incorporation at the gate oxide surface*, Appl. Phys. Letters, Vol. 66, pp. 3495-3497, 1995.

…

Chapter 5
Optimization of ultra-thin plasma nitrided oxides

5.1 Introduction

5.1.1 Motivation

In chapter 2, we have seen that the use of a plasma nitrided oxide in MOS transistors yields lower gate leakage current than pure oxide gate dielectric (factor ten reduction in leakage current at a given EOT). In our study, plasma nitrided oxides formed using the DPN process have been studied because of tool availability and because they yield good uniformity. In chapters 3 and 4, electrical and physical characterization techniques for such ultra-thin plasma nitrided oxides have been optimized. These characterization techniques will be used in this chapter with a view to optimize ultra-thin plasma nitrided oxides. The impetus for improvement has been to achieve low EOT with low gate leakage current density (J_G) while maintaining high effective carrier mobility.

5.1.2 Chapter Overview

Optimization of the base oxide, the plasma nitridation and the post nitridation anneal are presented. The results of the reliability aspects of the obtained ultra-thin plasma nitrided oxide is then summarized. An alternative to the conventional plasma nitridation process is then proposed for possible extension of plasma nitrided oxides. Finally, the ultra-thin plasma nitrided oxides optimized during this work are benchmark to conventional pure oxide and to latest high-K data. Both J_G-EOT and carrier mobility are included.

5.2 Description of the Decoupled Plasma Nitridation process

Plasma nitrided oxide films are formed in three steps (Figure 5.1):
First, a pure thin oxide film is grown using in-situ steam generation oxidation (ISSG) or rapid thermal oxidation (RTO). Note that prior to this first step, the wafer has been cleaned in an HF diluted solution. This surface preparation prior to the oxide growth is of great importance since the SiO_2/Si interface is a more significant part of the total film as it gets thinner.
The next step consists in exposing the oxide to a high density N_2-plasma for nitridation, which determines the nitrogen incorporation. When giving energy to an atom or a molecule, one or more electrons can escape the confinement to that atom or molecule. This phenomenon is called ionization and can occur by many methods such as DC current, ultraviolet radiation or radio frequency (RF) electric fields. This latter is the one that is used in this work for plasma nitridation. The plasma is formed when an avalanche of ionization occurs and is thus formed of ions, electrons and neutral species that can be highly reactive. Recombination can occur between an ion and an electron (the plasma is therefore emitting light). For a fixed power input, ionization and recombination are in equilibrium. In the case of nitridation, the inert gas used to form the plasma is N_2. He gas can be added to create more activated species. A schematic of the plasma nitridation chamber used is presented in Figure 5.2(a).
The final step of the DPN process is a high temperature anneal, so-called post nitridation anneal (PNA). The purpose and impact of this PNA on the SiON stack will be further detailed in this chapter.
There is no vacuum break between these three steps as they are performed in a cluster tool (in Figure 5.2(b)).

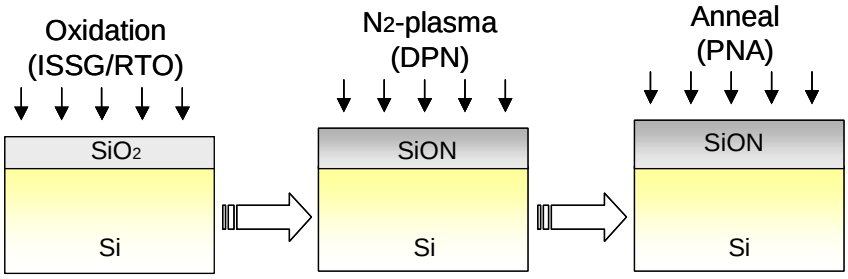
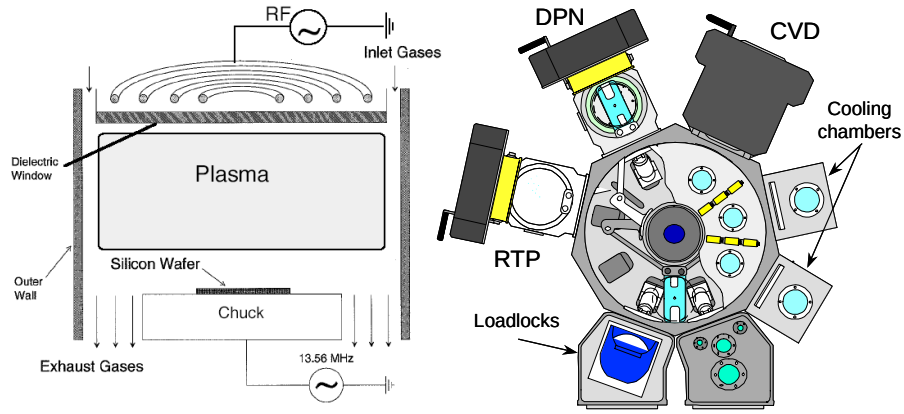


Figure 5.1: *Scheme of the DPN process. It consists of three steps: growth of a high quality oxide, plasma nitridation and a high temperature anneal.*



…… …. (a) (b)

Figure 5.2: *(a) Scheme of the DPN chamber. RF electric fields are used to create the plasma. The silicon wafer is located below the plasma. (b) Scheme of a gate dielectric single-wafer cluster tool showing the loading and unloading ports, an RTP, a plasma and a chemical vapor deposition chambers.*

5.3 Optimization of the base oxide

5.3.1 Comparison of the RTO and ISSG oxidation process

Growing thermal oxide on Si by exposing Si to O_2 at elevated temperatures is certainly one of the simplest processing steps in a MOS process flow. Moreover, thermal oxides consume Si during growth, thereby continuously creating a new and fresh interface. The simplicity of the process as well as the perfection of the resulting interface is largely responsible for the choice of Si as the substrate material for integrated circuits. In this work, we have studied two different oxidation processes: Rapid Thermal Oxidation (RTO) and In-Situ Steam Generation oxidation (ISSG), that are both performed in a rapid thermal processing (RTP) chamber. There are two main reasons for this choice. First the RTP oxide offers better absolute thickness control in the sub-2 nm range and greater processing temperature variety than an oxide grown in a furnace. Second, an RTP chamber is much smaller than a furnace and can be easily clustered with other chamber such as a cleaning, plasma or /and deposition chambers. An example of such cluster tool is depicted in Figure 5.2(b).

This clustering capability is of great importance when forming ultra-thin oxide based gate dielectric, as any surface contamination will affect the physical and electrical properties of the gate dielectric.

RTO is a dry oxidation following the chemical reaction:

$$Si + O_2 \rightarrow SiO_2$$

ISSG is a semi-wet oxidation following the chemical reactions:

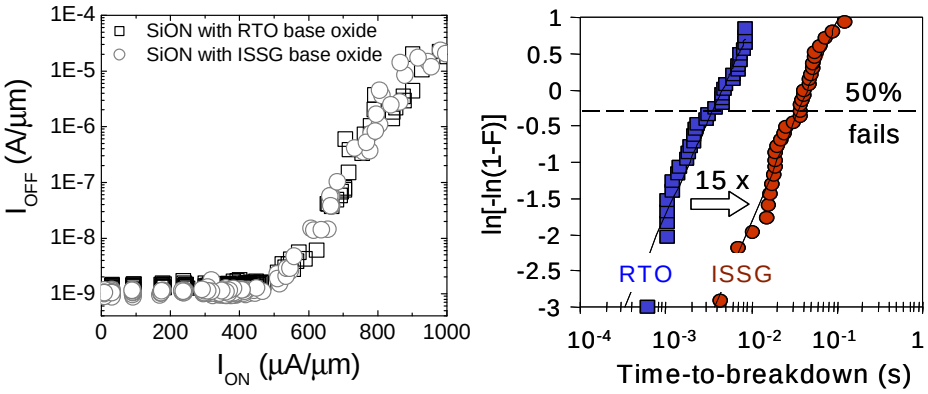
$$2H_2 + O_2 \rightarrow 2\,H_2O$$

$$Si + 2\,H_2O \rightarrow SiO_2 + 2H_2$$

Note that a full wet oxidation is not appropriate for growing thin oxide as H_2O as higher oxidation rate [1].

A comparison of the properties of RTO and ISSG oxides has been made. In Figure 5.3(a), the off-state versus the on-state drain current characteristics have been compared for NMOS transistors with a plasma nitrided oxide having an ISSG or RTO oxide. The EOT for both gate dielectrics is 1.4 nm. No major difference is observed between the two curves indicating that the electrical properties of NMOS transistors formed with a plasma nitrided oxide with an ISSG or RTO base oxide are similar. Comparable results were obtained for PMOS transistors.

The reliability of RTO and ISSG oxides has been also studied. In Figure 5.3(b), a comparison of the normalized time-to-breakdown Weibull distribution for RTO and ISSG oxides is presented. A factor 15 improvement in time-to-breakdown is measured for the ISSG oxide as compared to the RTO one. This could be attributed to the presence of H that might reduce the defect density in the oxide resulting in enhanced reliability [1], [3].



(a) (b)

Figure 5.3: *(a) Off-state (I_{OFF}) versus on-state (I_{ON}) drain currents for NMOS transistors having a plasma nitrided oxide formed with an RTO or ISSG oxide. The DPN and PNA conditions are fixed and the EOT is 1.4 nm for the two gate dielectrics. The width of these transistors is 10 μm and the supply voltage is 1 V. (b) Time-to-breakdown Weibull distributions for RTO and ISSG oxides.*

5.3.2 Scalability of the base oxide

The easiest way to reduce the thickness of plasma nitrided oxide films would be to scale the base oxide thickness.
SiON films formed with a base oxide of different thickness and with various N content have been characterized both physically and electrically. The physical thickness and N content, as measured with XPS, have been compared for SiON films having a 1 or a 2 nm thick base oxide (Figure 5.4(a)). First, it can be observed that increasing the amount of N yields an increase of the physical thickness of the films. There are several possibilities as to what caused this thickness increase:

1. One involves additional oxide formation from residual oxygen (O) in the DPN chamber. However, it is improbable that this parasitic re-oxidation account for the entire increase of the films thickness as a small amount of O was measured in the DPN chamber.
2. The second possible reason is that the Si substrate might get nitrided during the DPN treatment leading to the formation of a nitride-like interface. However, this hypothesis could not be verified by TEM cross section image where a nitride-like layer was not observed at the interface gate dielectric/Si substrate.
3. Another explanation could be related to the N incorporation mechanism itself where N atoms are exchanging with O atoms in the SiON matrix. We have shown previously (see chapter 4) that N atoms are preferably bonded to Si atoms in the SiON matrix. After the full nitridation process (including the PNA), some O and N atoms might not be bonded and might occupy interstitial sites. This may lead to volume change and to the observed increase in physical thickness.

Furthermore, it can be observed in Figure 5.4(a) that a faster increase of the physical thickness is measured for the samples with the 1 nm RTO base oxide as compared to the 2 nm one. This could be explained by the fact that the thinner the base oxide is, the higher the probability for highly reactive N species to reach the SiO_2/Si interface. The released O atoms might simply re-oxidize with the closed Si atoms of the substrate surface. This will result in a parasitic re-oxidation of the SiON film. This parasitic re-oxidation can seriously limit the down scaling of plasma nitrided films unless the energy of the species in the plasma is reduced so that N species do not reach the SiO_2/Si interface. In the next section, the optimization of the plasma with a view to reduce the energy of the N ions and neutrals is discussed.

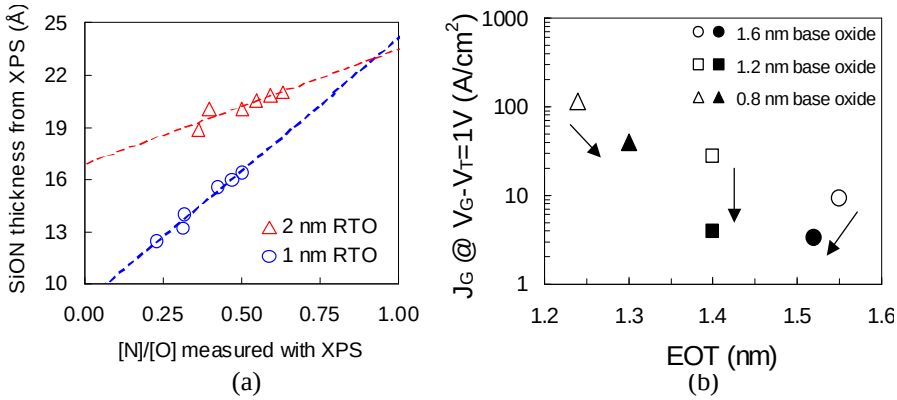


Figure 5.4: *(a) SiON thickness versus the [N1s]/[O1s] ratio as measured with XPS for plasma nitrided oxides with a base oxide thickness of 1 or 2 nm. Variation in the plasma nitridation process were performed to create various N content. (b) J_G as a function of EOT for NMOS transistors having a plasma nitrided gate oxide. The base oxide thickness and DPN time of the plasma nitrided oxide films are varied. The open and full symbols represent a small and a large amount of N in the SiON film, respectively.*

Another negative effect of thinning the base oxide in the sub-nm regime is that above a certain amount of N in the film, a saturation state is achieved prohibiting any further incorporation of N. This saturation phenomenon together with the large increase of the physical thickness have a strong impact on the electrical properties of the SiON films. The EOT and J_G of NMOS transistors with various SiON films have been compared. In Figure 5.4(b) a comparison of the trend J_G-EOT has been made for SiON films having a 1.6, 1.2 and 0.8 nm base oxide. The time of the DPN process was changed resulting in a variation of the N content in the film. The PNA step was kept constant for all the samples. For the samples with the thickest base oxide (1.6 nm) a decrease of both J_G and EOT is observed when increasing the amount of N in the film. This decrease in EOT is no more observed for the samples having a base oxide of 1.2 nm and even an increase of the EOT is observed for the samples with the thinnest base oxide (0.8 nm). As already discussed, increasing the amount of N results in a higher dielectric constant (ε) of the film and a thickening of the film. This physical thickening is more pronounced for SiON films having a thin base oxide. Therefore, while J_G is decreasing because of the thickening of the film, the amount of N incorporated in the film does not yield to a sufficient increase of ε to enable a decrease of the EOT (see definition of the EOT in Appendix A). In this experiment, a saturation of the amount of N in the SiON films having the thinnest base oxide has been also observed resulting in an unchanged value of ε. The EOT is thus increasing with the physical thickness.

Consequently, when thinning the base oxide thickness, the process window for gate dielectric optimization is dramatically reduced, resulting in a trade-off between scaling the EOT and reducing J_G.

Finally, because thinning the base oxide will increase the probability of having N located at the SiO_2/Si interface, higher defect density, such as interface states and fixed charges, will be created. As a consequence, the SiON films will exhibit a very poor reliability and a considerably large V_T shift will be observed for transistors having a plasma nitrided oxide formed with an ultra-thin base oxide [4].
As a conclusion to the above results, the thickness of the base oxide was not aggressively scaled in our experiment: it was always kept above 1 nm. However, research efforts were focused on the realization of a low energy ion plasma to limit the incorporation of N at the interface SiO_2/Si interface.

5.4 Plasma optimization

5.4.1 From continuous wave to pulsed RF source power

5.4.1.1 Plasma characterization

One mechanism for N incorporation from the plasma into the gate oxide is breaking Si-O bonds by energetic nitrogen species. Ideally, larger N concentration should be at the gate electrode/gate dielectric interface for efficient blocking of B penetration, reduction in gate leakage current and for minimal channel mobility degradation. However, this N profile cannot be maintained if the energy of the N ions and neutrals is too large, as the N species will penetrate through the gate oxide into the transistor channel. This will prohibit any further downscaling of the SiON film as shown in the previous section. Furthermore, this will create defects at the interface gate dielectric/Si substrate that will affect the reliability of the film as well as the channel mobility yielding a degradation of the transistor performance.
The distribution of electron energies in an equilibrium plasma is Maxwellian and is described by the electron temperature (kT_e), which is proportional to the mean ion kinetic energy (E_i) [5]. Therefore, reducing kT_e results in an improved plasma nitridation process. One method to lower kT_e is to use a pulsed RF source power (p-RF) instead of continuous wave power (CW). Pulsing the RF source power (turning on and off the RF source power at kHz frequencies) allows the fastest electrons in the plasma to diffuse to the wall of the chamber during the off cycle, thus leaving and cooling the plasma. The heavier ions are too slow to escape, and leave the ion density of the plasma unchanged.
A Langmuir probe was used to measure kT_e at 1 µs intervals in the N_2 plasma, resulting in the ability to observe the periodic kT_e variations due to the pulsed source power. The probe was positioned at about 4 cm above the center of the wafer in the plasma reactor. A comparison of kT_e and the ion density have been made for p-RF and CW generators over a wide range of effective power. The effective power has been calculated as the power of the generator times the time it is turned on. As an example, if the plasma is switched on only 50 % of the total time and the RF generator has a power of 1000 W, then the effective power is 500 W. The electron temperature and ion density are presented respectively in Figure 5.5(a) and (b) for a CW and a p-RF source power at various effective power. Note that for a CW plasma,

the effective power is equal to the total power. The net result of pulsing the source power is a plasma with a higher ion density and a reduced mean ion energy when compared to CW plasmas [6].

Because the wafer experiences many thousands of p-RF cycles during typical processing, an important metric is the average electron temperature ($\langle kT_e \rangle$). We calculate this parameter as in (5.1), where the average is calculated in the normal way for a periodic parameter.

$$\langle kT_e \rangle = \frac{1}{\tau}\int_0^{\tau} kT_e(t)dt \tag{5.1}$$

An example of a time-resolved average $\langle kT_e \rangle$ measurement is shown in Figure 5.6, where the periodic changes of the p-RF plasma are clearly observed.

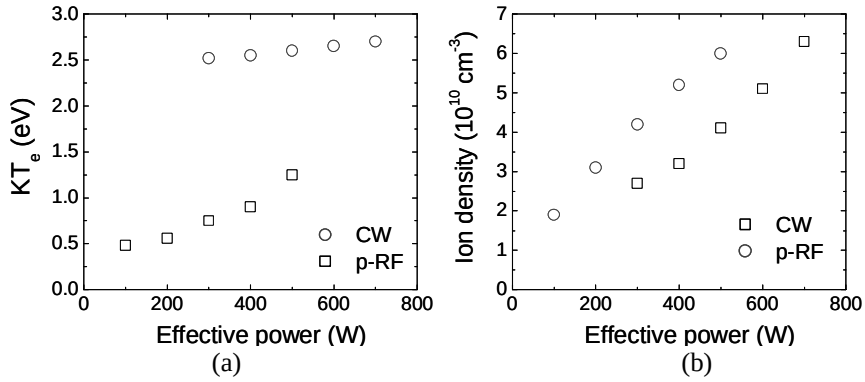


Figure 5.5: *(a) Electron temperature (kT_e) and (b) ion density as a function of effective power at 10 mTorr for CW and p-RF N_2 plasma at various effective power. The variation in p-RF kT_e at a given effective power results from the variation in duty cycle (fraction of the time where the p-RF plasma is switched on).*

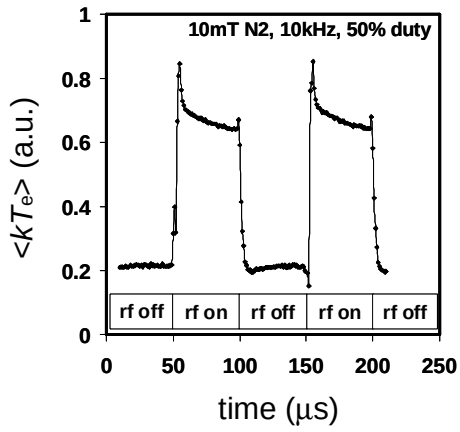


Figure 5.6: *Typical time-resolved electron temperature trace recorded with standard p-RF generator. Pulsing conditions are 10 kHz and 50 % duty cycle (i.e., the RF source power is on for 50 % of the pulsing period).*

5.4.1.2 Device performance

A comparison of the peak normalized transconductance (as a good indicator of the low field mobility) of both N-and PMOS transistor with plasma nitrided oxide formed with a p-RF or a CW plasma generator has been made at various N content [6]. For a given N content, devices with p-RF plasma processing show increased peak transconductance relative to devices with CW plasma (see Figure 5.7(a) and (b)).

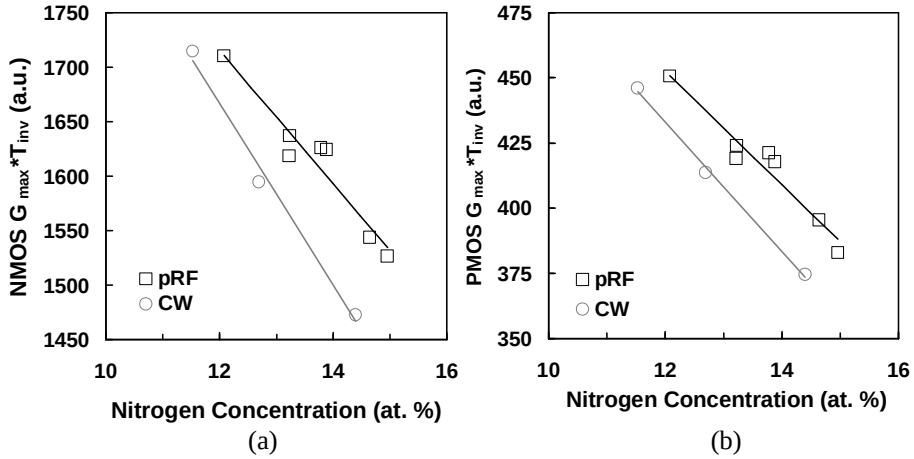


Figure 5.7: *Normalized peak transconductance for (a) NMOS and (b) PMOS having a plasma nitrided oxide formed with a CW or a p-RF source DPN at various N content. The channel length and width is*

respectively 3 and 10 μm [6].

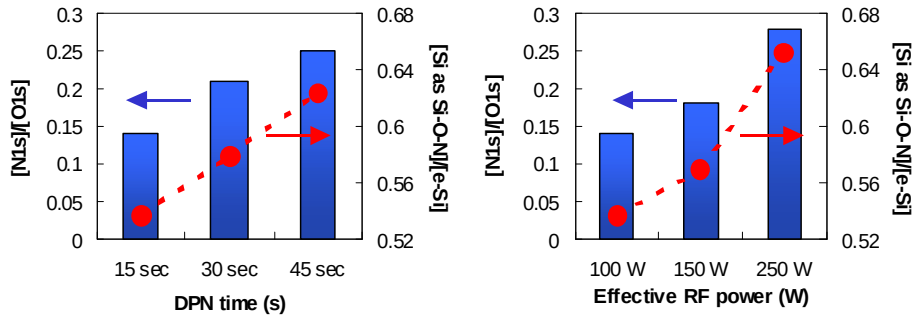
Finally, it was reported in the literature that transistors with a plasma nitrided oxide formed in a p-RF plasma exhibit reduced V_T shift and lower NBTI at all stress voltages and times [6].

5.4.2 Optimization of plasma generated by pulsed RF source power

Reducing the mean ion energy while maximizing the ion density has been the drive for further plasma optimization. In other words, the objective was to obtain a "soft" plasma in order to incorporate a large amount of N at the top surface of the dielectric. We have shown in section 5.4.1 that p-RF source power yields to plasmas with lower ion energy and higher ion density as compared to CW power. In this section, the p-RF source power has been further optimized by identifying key plasma parameters and by modifying the p-RF source itself [7]. The improved plasma process was then tested on N- and PMOS transistors.

5.4.2.1 Identification of key plasma parameters

Various plasma parameters have been investigated by performing Langmuir probe measurements and XPS analysis. The goal was to select the parameters that influence the most the characteristics of the plasma. The ratios [N1s]/[O1s] and [Si as SiON]/[elementary Si] ([e-Si]) have been measured by XPS as good indicators of the N content and the physical thickness of the SiON film, respectively. In Figure 5.8(a) and (b), the total nitridation time as well as the effective power have been varied. An increase of these two parameters results in an increase of both the N content and the physical thickness of the film. This is in accordance with Figure 5.4(a) showing that the incorporation of N leads to a physical thickening of the film.



(a) (b)

Figure 5.8: *[N1s]/[O1s] and [Si as SiON]/[e-Si] of DPN gate dielectric films formed with (a) various DPN time and (b) various plasma effective*

power.

Others parameters of the plasma such as the pulsing frequency and the pressure were studied. Their impact on the plasma properties was less significant. Still, it was observed that lowering the pressure as well as the pulsing frequency yields lower $< kT_e >$.

From these measurements, the effective power of the p-RF source as well as the time of the nitridation have been selected as key process parameters and have been thoroughly investigated with a view to obtain a low energy dense plasma. To compliment the Langmuir probe characterization of the p-RF plasmas, a retarding field analyzer (RFA) was used to measure the N ion energy distribution (IED) as a function of process and pulsing parameters. The RFA detector was mounted on the wafer surface at wafer center. Unlike the time-resolved kT_e measurements, the IED measurements here are time-averaged only. The RFA makes a direct measurement of the time-averaged ion population in the plasma at each energy. The effective power parameter can be dissociated in the duty cycle (DC), which represents the fraction of the time where the plasma is switched on, and in the peak power. The IED of these two variables have been measured. In Figure 5.9(a), the IED of three duty cycle conditions is presented:100, 50 and 10 %. Note that the 100 % duty cycle plasma is actually a CW plasma (i.e. not pulsed). In the p-RF IED profiles, there are two roughly Gaussian distributions observed, one at lower energy and one at higher, where the higher energy distribution is at about the same energy as the CW distribution. Decreasing the duty cycle yields a decrease of the high energy peak and an increase of the low energy peak. This result is in very good agreement with the desired plasma. Indeed, decreasing the high energy peak will make the plasma softer while the increase of the lower energy peak yields an increase of the N ion population.

A similar study was done for the peak power of the p-RF source. The N IED profiles of SiON films formed with plasmas at various p-RF peak powers are shown in Figure 5.9(b). The duty cycle was fixed at 10 %. Similar peaks as in Figure 5.9(a) are observed. Note that the high energy peak is already low because the duty cycle was considerably decreased. Increasing the p-RF source peak power results in an increase of the N ion population for both peaks. Increasing the peak power is therefore also beneficial to create low energetic dense plasma. However, high peak power yields to a rather unstable plasma and is therefore not recommended.

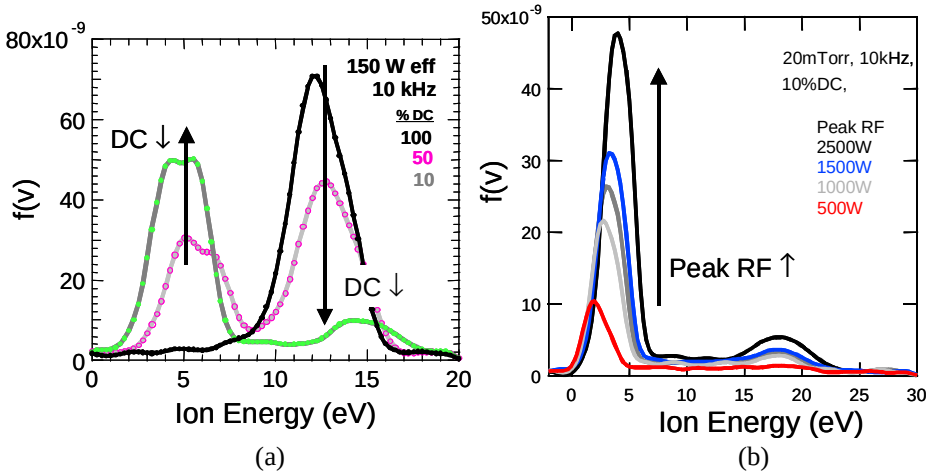


Figure 5.9: *(a) Nitrogen ions population distribution (proportional to f(V)) as a function of its energy at 10 mTorr, 10 kHz and at a fixed effective RF power of 150 W. The duty cycle (DC) was varied from 100 (CW) down to 10 %. (b) Nitrogen ions population as a function of its energy at 10 mTorr, 10 kHz and at a fixed DC of 10 %. The peak power of the plasma is varied from 500 up to 2500 W.*

As a conclusion, minimized $<kT_e>$, and therefore minimized average ion (electron) energy ($<E_{i(e)}>$) are achieved when operating at low duty cycle, that is to say at relatively long off-time per cycle. Using a low duty cycle plasma will also increase the ion (electron) density and therefore the amount of N incorporated in the oxide. For further increase of the N content, the time of the nitridation could be increased.

5.4.2.2 Towards "softer" plasmas

While extending the off-time per cycle reduces $<kT_e>$, the maximum measured electron temperature increases with increasing off-time per cycle, as shown in Figure 5.10(a). The maximum in kT_e, which is observed with the studied p-RF source has previously been observed for this type of pulsed plasma [7]. The increase of the maximum in kT_e can enable N atoms to be incorporated at the SiO_2/Si interface which is, as already discussed in this chapter, very harmful for the quality of the film and its scalability.

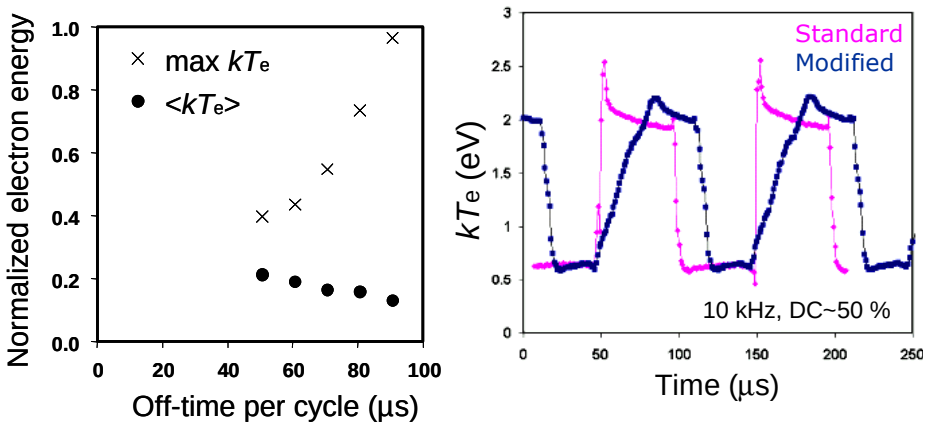
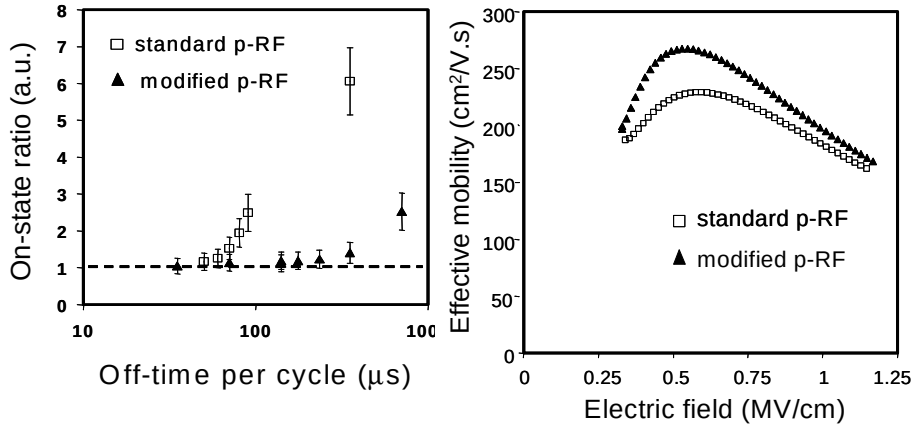


Figure 5.10: *(a) Maximum electron temperature (Max kT_e) and period averaged electron temperature (<kT_e>) for standard p-RF at 10 kHz and 10-50 % duty cycle as a function of the off-time per cycle (100-DC). (b) Comparison of the electron temperature for the modified and non-modified (standard) p-RF plasmas.*

Through optimization of the p-RF source plasma generator, it is possible to achieve stable plasmas with long off-time per cycle that do not exhibit significant kT_e overshoot during the RF-on portion of the cycle [8]. At this stage, the non-modified plasma will be called "standard" plasma and the modified one will be named "modified" plasma. The time-resolved electron temperature traces recorded with standard and modified p-RF generator are shown in Figure 5.10(b). It can be observed that the modified p-RF plasma does not exhibit kT_e overshoot during the very first µs of the on portion of the cycle. A metric for this on-state variation is the ratio of the maximum kT_e to the stable value of kT_e in the on-state; for example, the data in Figure 5.10(b) has an on-state ratio of approximately 1.3. An optimized p-RF process would have a ratio of 1.0 at all off-times per cycle that exhibit stable plasmas. Figure 5.11(a) is a plot of the on-state ratio for standard p-RF and modified p-RF hardware sets. The modified p-RF processes have on-state ratio near unity at off-time per cycle up to about 500 µs.



(a) (b)

Figure 5.11: *(a) On-state ratio, as defined in the text, for standard and modified p-RF. Data are in the ranges 1-50 kHz and 10-50 % duty cycle. (b) Low field effective channel mobility vs. vertical electric field for* L_G *= 3 µm and W=10 µm NMOS transistors with N dose of* 2×10^{15} cm^{-2}*. Both standard and modified p-RF were at 10 kHz and 30 % duty cycle. In both devices, CET was about 2.0 nm as determined by the RF impedance technique (see chapter 3).*

It was then interesting to study transistors with plasma nitrided oxides formed with this "softer" plasma.
Low field effective electron mobility measurements have been performed on NMOS transistors having a gate length and width of 3 and 10 µm, respectively. A comparison of the low field effective electron mobility of NMOS devices with plasma nitrided oxide formed using the standard and modified p-RF source was made (Figure 5.11(b)). Electron mobility is improved by 10 % at a field of 0.8 MV/cm. The improvements in carrier mobility here can be attributed to the improved control of the instantaneous electron temperature, as expected from Figure 5.11(a), and not to an absolute minimization of the high-energy ion population. Considering that in this experiment all splits were performed at 30 % duty cycle, further mobility improvement may be possible through the use of modified p-RF plasmas at even lower duty cycle.

5.5 Role and optimization of the Post Nitridation Anneal

In this section, the role of the Post Nitridation Anneal (PNA) in the formation of ultra-thin plasma nitrided oxides is studied. Different ambient have been studied with a view to optimize the plasma nitrided films.
First, XPS analyses have been performed on ultra-thin nitrided oxide that did not receive the final PNA step. The N1s and Si2p spectra measured with XPS were

studied. The N1s spectra is constituted of two peaks (Figure 5.12(a)): a large peak at a binding energy of 397.6 ± 0.1 eV and a smaller one at higher binding energy, 402.3 ± 0.1 eV. These binding energies correspond respectively to Si≡N and Si_2NO bonds as listed in Table 6 of chapter 4. The Si2p spectrum exhibits also two peaks consisting of a part due to elementary silicon (e-Si) and a part due to SiON (102-102.8 eV), as presented in Figure 5.12(b) and in chapter 4.

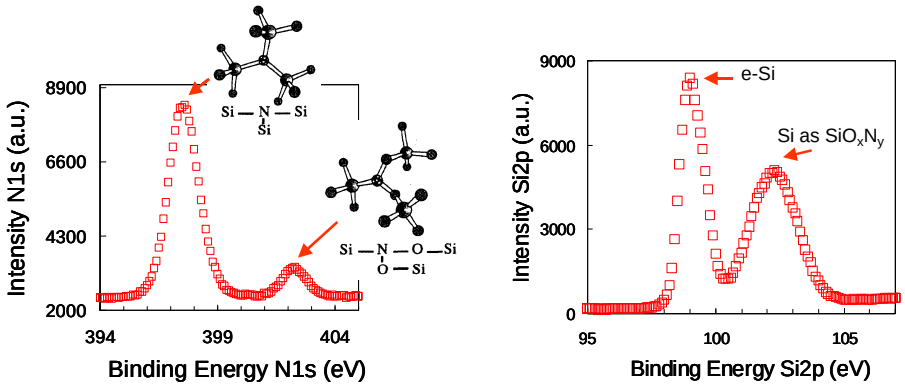


Figure 5.12: *(a) Spectrum of the N1s element after DPN process. Two peaks are measured with the XPS technique, showing the formation of mainly two types of bonds: Si_3N and Si_2NO. (b) Spectrum of the Si_2p peak are also shown. The elementary Si and the SiON peak can be observed.*

A possible explanation for the observed two peaks in the N1s spectrum is as follows. During the plasma nitridation step, (i.e. during the "soft" implantation of reactive N atoms in the thin SiO_2 film, as shown in Figure 5.13(a)), some N atoms are incorporated in the SiO_2 film. This will give rise to an unstable situation where some O atoms might be kicked-out of the SiO_2 matrix but, because there is no diffusion of O or N species (the plasma nitridation is performed at room temperature), most of the O and N atoms are still in the SiON film, although not in a stable position. A scheme of the SiON film formed without a PNA is presented in Figure 5.13(b).

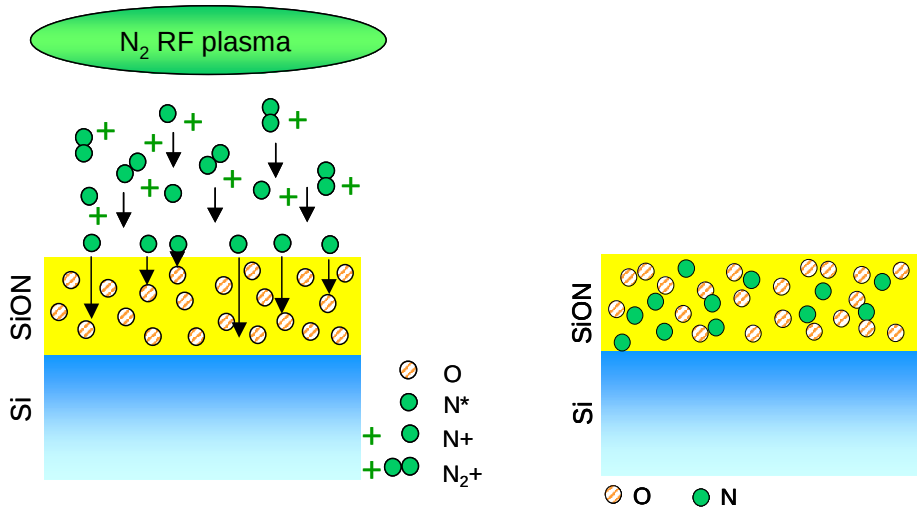


Figure 5.13: *(a) Scheme of the plasma nitridation process (DPN). First the pulsed RF plasma is generating ionic species N^+ and N^{2+} that diffuse towards the SiO_2 surface. Then, the ions are neutralized at the distance D of the SiO2 surface and give rise to highly reactive neutral N species (N*). Finally, N* species are implanted in the SiO_2 film (the incorporating depth depends on the N* initial kinetic energy). (b) Scheme of the SiON film after DPN. The "soft implantation" of N atoms results in a non equilibrium situation.*

The N1s and Si2p spectra were then measured on a similar SiON film except that the film was annealed. The N1s spectrum exhibits only one N1s peak indicating that there is mainly one type of N in the DPN oxynitride film, as presented in Figure 5.14(a). The binding energy measured for the N1s peak is 397.6 ± 0.1 eV. As presented in chapter 4, this indicates that N atoms are mainly bonded to Si atoms creating Si≡N bonds.

During the high temperature anneal (PNA) that follows the DPN processing step, the O and N atoms diffuse and react, giving rise to different bonding configurations. The probability of having an atom of O or N diffusing towards the top surface or the interface is estimated as equal. Therefore, some O in the form of O_2 and some N in the form of N_2 can outgas from the SiON layer. Furthermore, some N atoms can exchange with O atoms in the SiO_2 matrix to form extra Si≡N bonds. Finally, some O atoms can diffuse towards the SiO_2/Si interface and re-oxidize this interface, as mentioned earlier in this chapter. These various possible mechanisms are presented in Figure 5.14(b). The role of the PNA is therefore to stabilize the atoms in the SiON matrix.

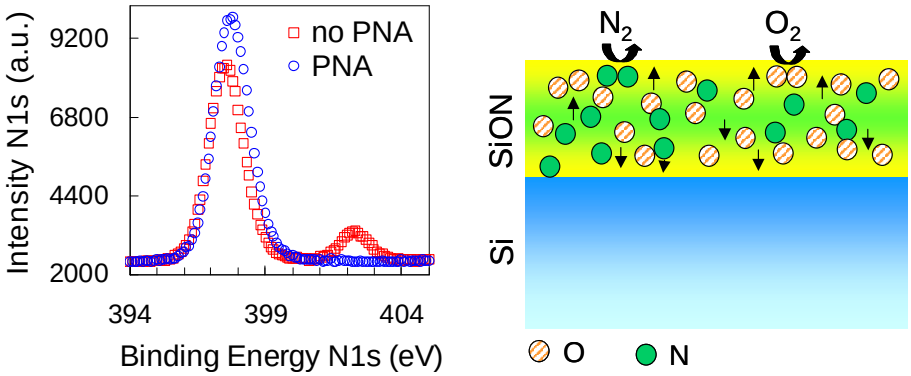


Figure 5.14: *(a) Spectra of the N1s element after DPN and DPN+PNA processes. After the PNA step (1000 °C, 15 s, O_2), only one peak is measured at a binding energy close to N as Si_3N_4, indicating that most of the N is binding to Si atoms after the high temperature anneal (PNA). (b) Scheme showing various possible O and N diffusion mechanisms that could occur in the SiON layer during the PNA step.*

At this stage, it is interesting to study different ambient for the PNA. Three ambient were investigated: two in an inert ambient (in He and in N_2) and one in an oxidizing ambient (in O_2). A sample without PNA was also compared to the samples with the various PNA ambient. This experiment was performed on thin (1 nm) and "thick" (2 nm) base oxides (RTO) from which similar results were obtained. The same DPN step was processed on all the samples. The temperature and time of this anneal was set to 1000 °C and 15 s respectively. The N1s and O1s spectra have been measured with XPS and have been plotted in Figure 5.15(a) and (b), respectively. It is interesting to notice that only one N1s peak is observed for all samples at the binding energy of 397.6 ± 0.1 eV, indicating that N atoms are mainly bonded to Si atoms (Si≡N type of bonds) and this independently of the PNA ambient. Similarly, the peak of the various O1s spectra is located at a binding energy of 532.3 ± 0.1 eV, indicating that O is mainly bonded to Si atoms and not to N ones and this independently of the PNA ambient.

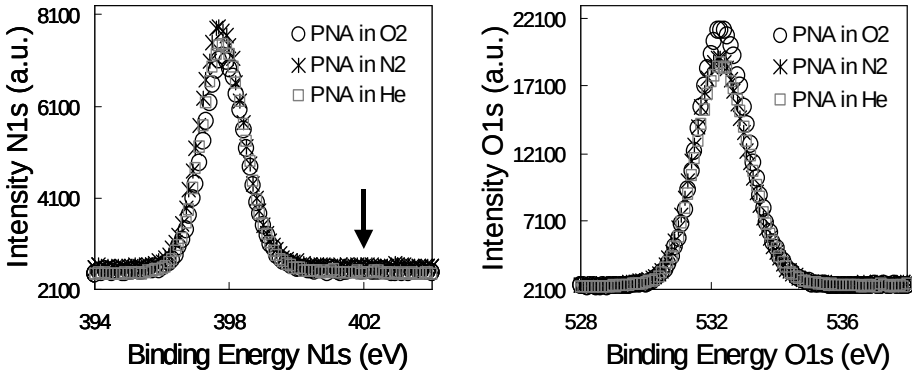


Figure 5.15 : *(a) N1s and (b) O1s spectra of ultra-thin SiON films formed with a PNA in O_2, N_2 or He. Only one N1s peak (Si≡N bonds) is observed even after a PNA in an oxidizing ambient.*

To further understand the kinetic mechanism during these PNA, the ratio [N1s]/[O1s] and the physical thickness were calculated from the XPS analyses. The results are presented in Figure 5.16.

For SiON films having received the PNA anneal in an inert ambient (namely N_2 or He), the ratio [N1s]/[O1s] is higher than for the SiON films having received a PNA in O_2 or no PNA treatment. This result is in accordance with the previous observations on the N incorporation mechanism where N atoms exchange with O atoms in the SiON film (see chapter 4) Moreover, it can be observed in Figure 5.16 that the physical thickness of the samples annealed in an inert ambient is similar to the one of the sample without PNA. This indicates that no extra N or O atoms are incorporated in the film during the PNA and that the O atoms are outgasing from the film rather than diffusing towards the interface to form a sub-oxide.

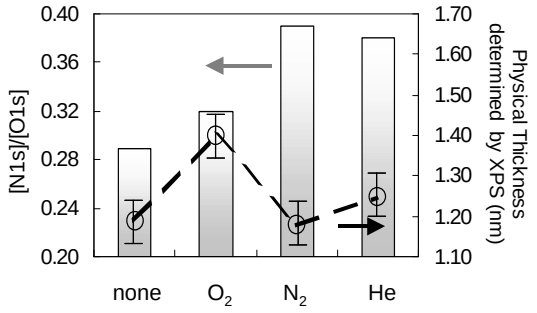


Figure 5.16: *Comparison of the ratio [N1s]/[O1s] and the physical thickness, as determined by XPS, for SiON films having a PNA treatment*

in various ambient. A sample without PNA is also shown.

For SiON films having received the PNA anneal in an oxidizing ambient (namely O_2), the [N1s]/[O1s] ratio is sensibly lower than the one of the samples annealed in an inert ambient (Figure 5.16). This can be explained by the fact that during the PNA in O_2 some extra O atoms are incorporated in the film resulting in a decrease of the ratio [N1s]/[O1s]. It was observed from Figure 5.15(b), that the O atoms were bonded mainly to Si atoms even after a PNA in O_2. Note that a higher binding energy was measured for the Si2p as SiON peak, when performing the PNA in O_2. The extra O atoms, incorporated during the PNA, seem therefore to diffuse towards the interface and to re-oxidize it. This is also confirmed by the increase in physical thickness for the SiON film annealed in O_2, as shown in Figure 5.16. Such a process can yield better channel mobility and reliability of the SiON film, as it has an oxide like interface with the Si channel, but can also be a showstopper for further downscaling.

Finally, it is interesting to have a closer look at the measurements obtained for the SiON sample that did not receive the final PNA step. While the physical thickness of the film is rather similar to the one of the samples annealed in an inert ambient, the ratio [N1s]/[O1s] is sensibly lower. This could be attributed to a slow re-oxidation of the N located at the top surface of the SiON film in the air of the cleanroom where the samples were stored. Indeed, as mentioned above, after DPN, the N atoms are not solidly bonded to Si atoms. Therefore a slow re-oxidation of nitrogen in the air could occur forming volatile compounds such as NO and N_2O. Consequently, clustering the PNA step with the DPN plasma processing is a must in the control of the thickness and stochiometry of thin SiON films.

As a conclusion, the PNA is an important step in the fabrication of ultra-thin plasma nitrided films and should be clustered with the DPN process. Its role is to stabilize the SiON film after the plasma nitridation step. It can also healed some defects created during the plasma step, as it is done at high temperature. The ambient (oxidizing or inert) has an impact on the composition and thickness of the final SiON film.

5.6 Extendibility of plasma nitrided gate oxides

As described in the introduction of this thesis, the thickness of the gate dielectric is supposed to scale linearly with the channel gate length of the MOS transistors. We have seen that this decrease in thickness is not straightforward because of the stringent gate leakage current requirement (see chapter 2). Process wise, we have also seen in this chapter that further downscaling of plasma nitrided oxides is a real challenge. Gate dielectrics with a higher dielectric constant (so called high-*K* dielectrics) are of course very attractive since low J_G and EOT in the nanometer range can be achieved. A lot of research efforts have been focused in the past years

on the investigation of many high-*K* materials for the potential replacement of SiON dielectrics ([22], [23], [24], [25] and [26]). However, there are still many issues to be solved as already mentioned in chapter 2. Furthermore, there is resistance in introducing high-*K* material in a production fab due to contamination risks. With this perspective, it is important to evaluate possible extensions of SiON films.

In the previous sections of this chapter, we have seen limitations in the downscaling of the plasma nitrided oxides because of saturation of N in the layer. It was also shown that N is mainly bonded to Si atoms in the SiON matrix. A possible extension to plasma nitrided oxide could be to form an oxide with an excess of Si in it, so called a silicon rich oxide (SRO). The standard DPN nitridation process could be then used on the SRO film to obtain a nitrided oxide. More N atoms could be incorporated in the film enabling further downscaling of SiON gate dielectrics. This idea is currently being investigated and only first results will be shown in this thesis.

5.6.1 Formation of the Silicon Rich Oxide layer

The SRO film is formed in a single wafer rapid thermal chemical vapor deposition (RTCVD) system. SiH_4 and N_2O are applied as reactants, N_2 as carrier gas. In order to obtain SiON films with small EOTs, the SRO film must be reduced; thereby very low deposition rate must be achieved. An HF clean is also used before the SRO process to minimize the initial oxide layer for better EOT scaling. Table 5.1 summarizes the process conditions used for the first experiments. To achieve a low deposition rate, depositions were carried out at 700 °C heater temperature, equivalent with a wafer temperature of approximately 650 °C and at the pressure of 50 Torr. The deposition rate obtained was 0.084 nm/min which showing that thin SRO films can be deposited.

Temperature (^oC)	Pressure (Torr)	SiH_4/N_2O	Deposition rate (nm/min)
700	50	0.2	0.084

Table 5.1: SRO deposition conditions.

With these process conditions, the thickness and uniformity of the SRO layers have been investigated at various deposition times. An initial thickness of 0.5 nm can be observed before any SRO deposition (Figure 5.17(a)). This can be explained by the fact that in the process recipe a pre-deposition step is included to stabilize the gas flows (SiH_4 and N_2O gas). The 0.5 nm initial oxide is due to the surface oxidation by N_2O in this pre-deposition step.

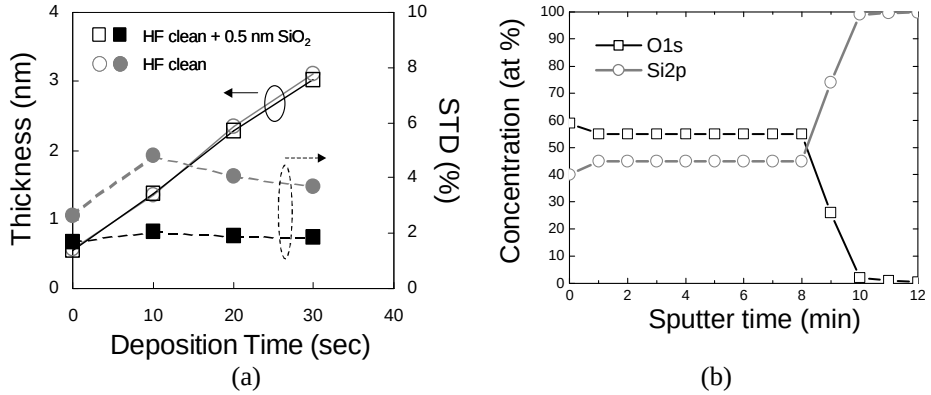


Figure 5.17: *(a) Thickness (left axis, open symbols) and its standard deviation, STD, (right axis, closed symbols) of SRO layers deposited at various times. Two surface preparations are compared: HF clean (circles) and HF clean with a 0.5 nm SiO_2 (squares). (b) Stochiometry of SiO_2 and SRO layers, as measured by XPS.*

The standard deviation calculated from thickness measurements within the wafer is used to evaluate the uniformity of the SRO layers within the wafer. The standard deviation is plotted on the second y-axis of Figure 5.17(a). It can be observed that the uniformity of the SRO layers directly deposited after the HF clean treatment is rather poor and especially not stable with the deposition time. In order to improve the uniformity, an ultra-thin oxide of 0.5 nm was grown after an HF clean treatment and before the SRO deposition. As can be seen in Figure 5.17(a), the SRO films deposited on 0.5 nm SiO_2, have the same thickness but an improved within wafer uniformity. It indicates that N_2O does not oxidize the surface further if there is already a 0.5 nm oxide present.
The stochiometry of the SRO layer has been measured by XPS analysis. A comparison of the stochiometry of a SiO_2 and a SRO film of same thickness has been made and is shown in Figure 5.17(b). The results indicate the SRO film contains about 20% excess Si compared to stoichiometric SiO_2.

5.6.2 Formation and characterization of plasma nitrided Silicon Rich Oxide gate dielectric

DPN nitridation has been performed on the deposited SRO film. A PNA was also performed after the DPN step. The amount of N incorporated in 1.4 nm SRO based and 1.4 nm SiO_2 based (in this experiment: ISSG oxide) SiON insulators have been compared for various nitridation conditions. The ratio [N1s]/[O1s] has been measured by XPS for the two types of films, as presented in Figure 5.18. The difference between the measured [N1s]/[O1s] ratio of SRO based and SiO_2 based SiON films is also shown. A net higher amount of N has been measured for the SRO

based SiON relative to SiO_2 based SiON. The SRO based SiON films contains more than N than the SiO_2 based SiON films for all the plasma nitridation conditions studied. This result confirms that more N atoms are incorporated in SRO than in SiO_2 oxides and also validates the incorporation mechanism of N in the oxide for plasma nitridation process.

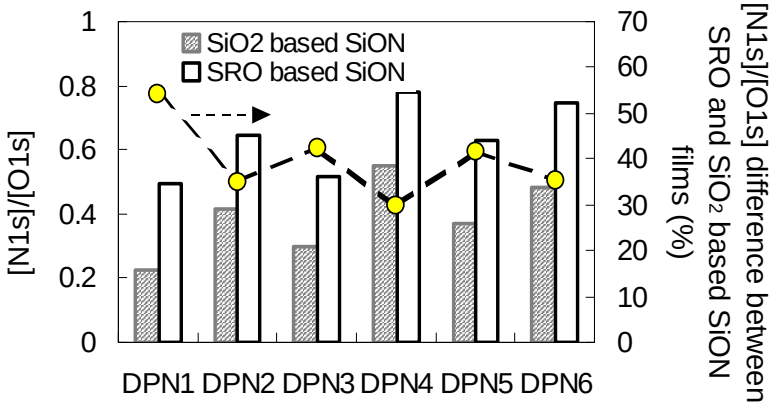


Figure 5.18: *[N1s]/[O1s] ratio measured by XPS for SRO based and SiO_2 based SiON films. Various DPN treatments have been applied on 1.4 nm SRO and 1.4 nm SiO_2 oxides yielding various N concentrations. The difference between the two ratios is plotted on the second y axis.*

5.7 Conclusions

Ultra-thin plasma nitrided oxides have been optimized with the objective to decrease J_G and maximize carrier mobility. It was found that while the base oxide cannot be aggressively scaled, plasma optimization yields better mobility thereby increase transistor performance. A summary of the EOT versus gate leakage current density of NMOS devices with plasma nitrided oxides is shown in Figure 5.19. EOT down to 1.2 nm has been achieved with a gate leakage current density of 40 A/cm^2 at 1 V operating voltage.

A thorough study of the role of the post nitridation anneal step confirmed the N incorporation mechanism in the SiO_2 matrix, as already proposed in chapter 4. Because N is mainly bonded to Si atoms in the SiON film, an alternative to the classical oxide based plasma nitridation process has been proposed. Silicon rich oxide (SRO) as a replacement to pure oxide prior to the plasma nitridation process results in lower gate leakage current at a given EOT because of the enhancement of N in the film. More than 30 % of N was found in the silicon rich oxide based SiON relative to SiO_2 based SiON gate dielectrics.

The reliability study of the plasma nitrided oxides was not part of this thesis work. Yet, the reliability of the optimized plasma nitrided oxides presented above has been investigated within the research team. It was found that the intrinsic reliability of these ultra-thin films is strongly dependent on the extrapolation law to operating voltage ([17], [12]). Furthermore, the present reliability specification assuming that the first gate oxide breakdown is fatal for the entire circuit is currently under debate ([14], [15] and [16]).

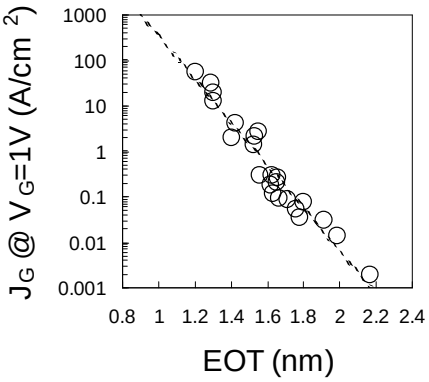


Figure 5.19: *Summary of gate leakage current densities and EOTs obtained for various plasma nitrided oxides studied in this work.*

5.8 References

[1] B.E. Deal and A.S. Grove, *General relationship for the thermal oxidation of silicon*, Journal of Applied Physics, Vol. 36, pp. 3770-3778, 1965.

[2] S. Bruyere, F. Guyader, W.D. Coster, E. Vincent, M. Saadeddine, N. Revil and G. Ghibaudo, *Wet and dry ultrathin oxides. Impact on gate oxide and device reliability*, Microelectron. Reliability, Vol. 40, pp. 691, 2000.

[3] Y. Murakami, T. Shiota, T. shingyouji and H. Abe, *Effect of oxidation ambient on the dielectric breakdown characteristics of thermal oxide films of silicon*, Journal of Applied Physics, Vol. 75, pp. 5302-, 1994.

[4] A. Veloso, F. N. Cubaynes, A. Rothschild, S. Mertens, R. Degraeve1, R. O'Connor, C. Olsen, L. Date, M. Schaekers, C. Dachs, M. Jurczak, *Ultra-thin Oxynitride Gate Dielectrics by Pulsed-RF DPN for 65 nm General Purpose CMOS Applications*, ESSDERC 2003, pp. 239-242

[5] M. Liberman and P. Lichtenberg, *Principles of Plasma Discharges and Materials Processing,* pp. 306-307, John Wiley& Sons, 1994.

[6] P.A. Kraus, K. Ahmed, T.C. Chua, M. Ershov, H. Karbasi2, C.S. Olsen, F. Nouri, J. Holland, R. Zhao, G. Miner and A. Lepert, *Low-energy Nitrogen Plasmas for 65-nm node Oxynitride Gate Dielectrics: A Correlation of Plasma Characteristics and Device Parameters*, Symp. On VLSI Tech. Digest, pp. 143-144, 2003.

[7] S. Ashida, M. R. Shim, and M. A. Lieberman, *???*, J. Vac. Sci. Technol., **A14** (1996) 391.

[8] P.A. Kraus, T.C. Chua, A. Rotschild, F.N. Cubaynes, A. Veloso, S. Mertens, L. Date, T.M. Bauer, K.Z. Ahmed, J. Campbell, F. Nouri, J. Cruse, R. schreutelkamp and M. Schaekers, *Further optimization of plasma nitridation of ultra-thin oxides for 65 nm node MOSFETs*, to be presented at the ECS conference in ???, 2004.

[9] R. Degraeve, G. Groeseneken, R. Bellens, M. Depas and H.E. Maes, *A consistent model for the thickness dependence of intrinsic breakdown in ultra-thin oxides,* IEDM Technical Digest, pp. 863-866, 1995.

[10] J.H. Stathis, *Percolation models for gate oxide breakdown,* Journal of Applied Physics, Vol. 86, p. 5757-5766, 1999.

[11] R. Degraeve et al., *New insights in the relation between electron trap generation and the statistical properties of oxide breakdown,* IEEE Trans. Elect. Dev., Vol. 45, pp.904-911, 1998.

[12] R. Degraeve, J.L. Ogier, R. Bellens, P.J. Roussel, G. Groeseneken and H. E. Maes, *A New Model for the Field Dependence of Intrinsic and Extrinsic Time-Dependent Dielectric Breakdown,* IEEE Trans. Elect. Dev., Vol. 45, pp. 472-481, 1998.

[13] P.J Roussel, R. Degraeve, G.V. van den Bosch, B. Kaczer and G. Groeseneken, *Accurate and robust noise-based trigger algorithm for soft breakdown detection in ultrathin gate dielectrics*, IEEE Trans. Device and Materials Reliability, Vol. 1, pp. 120-127, 2001.

[14] B. Kaczer, R. Degraeve, M. Rasras, K. van de Mieroop, P.J. Roussel and G. Groeseneken, *Impact of MOSFET gate oxide breakdown on digital circuit operation and reliability*, IEEE Trans. Electron Devices, Vol. 49, pp. 500-506, 2002.

[15] B. Kaczer and G. Groeseneken, *Potential vulnerability of dynamic CMOS logic to soft gate oxide breakdown*, IEEE Electron Device Lett., Vol. 24, pp. 742-744, 2003.

[16] B.E. Weir, M.A. Alam, J.D. Bude, P.J. Silverman, A. Ghetti, F. Baumann, P. Diodato, D. Monroe, T. Sorsch, G.L. Timp, Y. Ma, M.M. Brown, A. Hamad, D. Hwang and P. Mason, *Gate oxide reliability projection to the sub-2 nm regime*, Semicond. Sci. Technol, Vol. 15, pp. 455-461, 2000.

[17] E.Y. Wu, E. Nowak, L.K. Han, D. Dufresne and W.W. Abadeer, *Nonlinear Characteristics of Weibull Breakdown Distributions and Its Impact on Reliability Projection for Ultra-Thin Oxides*, IEDM Technical Digest, pp. 441-444, 1999.

[18] R. Degraeve, B. Kaczer, F. Schuler, M. Lorenzini, D. Wellekens, P. Hendrickx, J. Van Houdt, L. Haspeslagh, G. Tempel and G. Groeseneken, *Statistical model for Stress-Induced Leakage Current and pre-breakdown current jumps in ultra-thin oxide layers*, IEDM Technical Digest, pp. ???, 2001.

[19] G. Cellere, L. Larcher, M.G. Valentini and A. Paccagnella, *Micro breakdown in small-area ultrathin gate oxides*, IEEE Trans. Electron Devices, Vol. 49, pp. 1367-1374, 2002.

[20] B.P. Linder, S. Lombardo, J.H. Stathis, A. Vayshenker and D.J. Frank, *Voltage dependence of hard breakdown growth and the reliability implication in thin dielectrics*, IEEE Electron Devices Lett., Vol.23, pp. 661-663, 2002.

[21] M.A. Alam, J. Bude and Andrea Ghetti, *Field Acceleration For Oxide Breakdown - Can An Accurate Anode Hole Injection Model Resolve the E vs. 1/E Controversy ?*, Proc. IRPS, pp. 21-26, 2000.

[22] H.S. Kim, S.A. Campbell and D.C. Gilmer, *Charge trapping and degradation in high-permittivity TiO_2 dielectric films,* IEEE Electron Dev. Lett., Vol. 18, pp.465-467, 1997.

[23] M. Houssa, R. Degraeve, P.W. Mertens, M.M. Heyns, J.S. Jeon,A. Halliyal and B. ogle, *Electrical properties of thin SiON/Ta2O5 gate dielectric stacks*, Journal of Applied Physics, Vol. 86, pp. 6462, 1999.

[24] J. Morais, E.B.O. da Rosa, L. Miotti, R.P. Pezzi, I.J.R. Baumvol, A.L.P. Rotondaro, M.J. Bevan and L. Colombo, *Stability of zirconium silicate films on Si under vacuum and O_2 annealing,* Applied Physics Letters, Vol. 78, pp. 2446, 2001.

[25] G.D. Wilk, R.M. Wallace, J.M. Anthony, *High-κ gate dielectrics: Current status and materials properties considerations*, journal of Applied Physics, Vol. 89, pp. 5243, 2001.

[26] G. Lucovsky, *Transition from thermally grown gate dielectrics to deposited gate dielectrics for advanced silicon devices: A classification scheme based on bond ionicity*, Journal of Vaccuum Science and Technology A, Vol. 19, pp. 1353-1561, 2001.

Chapter 6
Integration of ultra-thin plasma nitrided oxide in advanced MOS transistors

6.1 Introduction

6.1.1 Motivation

In the previous chapter, plasma nitrided gate dielectrics have been optimized to obtain the best trade off between low gate leakage current and high performance (EOT scaling with high channel mobility). In addition to the scaling of the gate dielectric, the oxide capacitance can be further increased by maximizing the polysilicon gate activation at the polysilicon/gate dielectric interface, thus reducing polysilicon depletion effects. Maximizing the activation in the polysilicon gate requires a high temperature anneal. However, the formation of abrupt ultra-shallow junctions, necessary to control SCE, requires minimum thermal budget for this anneal. There is therefore a compromise between getting a highly activated polysilicon gate and scaling the junctions.
The aim of this chapter is to optimize the gate stack yielding high gate activation and low polysilicon depletion while still being compatible with the formation of abrupt ultra-shallow junctions.

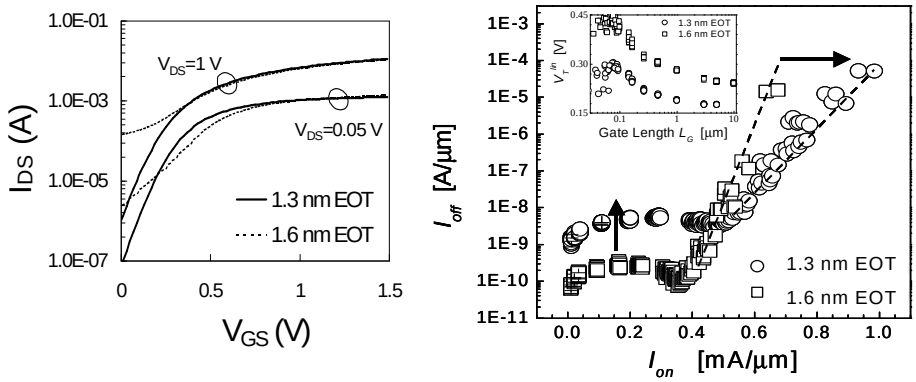
6.1.2 Chapter overview

This chapter will start with a short summary showing the impact of scaled plasma nitrided oxide on short channel transistor behavior. The optimization of the polysilicon gate is then detailed. First, the impact of the polysilicon gate activation on NMOS transistor behavior is addressed. The influence of the gate morphology on the

dopants activation and deactivation is then studied. Finally technological changes to maximize the gate activation at the gate/gate dielectric interface are proposed.
The compatibility of the optimized polysilicon gate stack with the formation of ultra-shallow junctions is then investigated. The impact of new techniques to form ultra-shallow junctions (namely pre-amorphized junctions and solid phase epitaxial regrowth junctions) on the polysilicon gate stack will be discussed.

6.2 Impact of downscaling the gate dielectric on short channel transistor performance

The optimization of ultra-thin plasma nitrided oxides for short channel CMOS transistors has been thoroughly studied in the previous chapter of this thesis. In this section, the characteristics of short channel transistors having an ultra-thin plasma nitrided gate oxide are studied. A comparison of the electrical behavior of NMOS transistors having an optimized plasma nitrided oxide of 1.6 or 1.3 nm EOT is shown. Figure 6.1(a) shows the linear and saturated drain current behavior of NMOS transistors having a channel length (L_G) of 45 nm and a 1.6 or 1.3 nm EOT plasma nitrided gate oxide. A better control of the gate on the channel is observed for the transistor with the thinnest gate dielectric. Indeed, as already mentioned in the main introduction of this thesis, scaling the gate dielectric thickness increases the induced channel charge at a given voltage.



(a) (b)

Figure 6.1: *(a) Linear and saturated drain currents for 45 nm NMOS transistors having a gate dielectric of 1.6 or 1.3 nm EOT. (b) I_{OFF}-I_{ON} characteristics of NMOS transistors with various channel lengths (width is 10 µm) and with two plasma nitrided gate oxide thicknesses: 1.6 and 1.3 nm EOT. The linear V_T is plotted in the inset of this figure for the same transistors.*

Furthermore, the combination of scaling the gate dielectric thickness together with the channel length acts to increase the drive current of the transistors, as illustrated in Equation 1.2 (chapter 1) and in Figure 6.1(b).
In Figure 6.1(b), the off-state versus on-state drain currents (I_{OFF} and I_{ON} , respectively) are plotted for NMOS transistors having two different plasma nitrided gate dielectrics (1.3 and 1.6 nm EOT) and having various channel length ranging from 10 μm down to 30 nm.
For short channel transistors (L_G < 0.13 μm) and at a given I_{OFF} ($I_{OFF} \geq 10^{-8}$ *A/μm*), a steep I_{OFF}-I_{ON} slope is observed for the NMOS devices with the thick plasma nitrided oxide (1.6 nm). This I_{OFF}-I_{ON} slope is less steep for devices with the thin plasma nitrided oxide (1.3 nm). This indicates that better SCE control is achieved when using a thin gate dielectric. Short channel devices (L_G=40 nm in this case) with a 1.3 nm plasma nitride oxide exhibit very good on-state performance with limited off-state leakage current.
For long channel devices ($L_G \geq$ 0.13 μm), I_{OFF} is not dominated by the gate leakage current but by V_T even for the transistors having a 1.3 nm EOT gate dielectric. Indeed, it can be observed in Figure 6.1(b), that for long channel devices, I_{OFF} is increasing and then decreasing, replicating the shape of the long channel V_T behavior, as depicted in the inset of Figure 6.1(b). The increase of V_T with decreasing the channel length (also called V_T roll-up) results from the influence of the halos dopants in the channel, which decreases the inversion region (associated with the increase in V_T). The difference observed in V_T for the NMOS devices with a 1.6 nm and 1.3 nm can be explained by the difference in thickness and nitrogen content.

As a conclusion, higher has been obtained for short channel transistors performance when scaling the plasma nitrided oxide and this, without increasing dramatically the transistor leakage current. This result underlines the successful integration of the ultra-thin plasma nitrided gate dielectrics optimized during this work (see previous chapter).

6.3 Polysilicon gate electrode engineering

6.3.1 Motivation

While decreasing the thickness of the gate dielectric, that is to say increasing the oxide capacitance (C_{ox}), yields higher transistor performance and allows a better control of the channel by the gate, the use of ultra-thin gate dielectric results in several undesired effects:

1. Gate leakage current: increases exponentially when scaling the gate oxide thickness (see chapter 2);

2. Carrier quantization in the channel due to quantum confinement ([1], [2]): adds about 0.4 nm to the total electrical thickness (leads to a decrease of the total inversion capacitance (C_{total}) see Figure 6.2);
3. Dopant diffusion from the gate polysilicon into the channel (especially in PMOS transistors): results in a shift of the threshold voltage and several other unwanted phenomena (V_T spread across the wafer, poor controllability of V_T and hence very tight process window, stronger V_T mismatch, etc…);
4. Carrier depletion in the polysilicon gate: yields a drop of the gate voltage (decrease of C_{total}, see Figure 6.2) because of a high electric field (as a result of the combination of high supply voltage and thin gate dielectric).

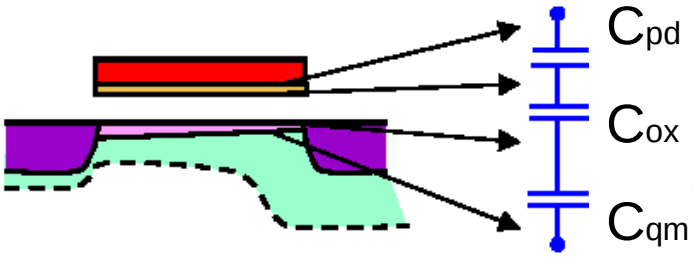


Figure 6.2: *Scheme of a MOS structure biased in inversion. The total inversion capacitance (C_{total}) is the sum of the oxide capacitance (C_{ox}) and the extra capacitances due to quantum confinement (C_{qm}) and polysilicon depletion (C_{pd}).*

The gate leakage current has been significantly reduced with the use of plasma nitrided oxides (see chapter 2 and 5). The other undesired effects mentioned above can be minimized by carefully optimizing the gate electrode and plasma nitrided oxide. CMOS processes down to, at least, the 65 nm CMOS technology node use polycrystalline silicon (polysilicon) as gate material. One of the main reasons is that polysilicon allows for the integration of "dual-flavored" gates: p-type and n-type polysilicon for PMOS and NMOS, respectively. This is important for having both types of transistors "surface channel" type to reduce short channel effects and to maximize drive current.

The gate depletion level is determined by different physical effects: dopant diffusion, dopant activation, dopant deactivation and penetration of the dopant through the gate oxide (B penetration for PMOS transistors). Achieving good gate activation by having a sufficient amount of dopants at the polysilicon gate-oxide interface is a critical step in the CMOS fabrication process. Indeed, for process simplicity and therefore for lower cost, the gate is doped and activated using the same ion implantation and annealing steps that form the source and drain. This results in a compromise between a fully activated polysilicon gate to assure an ideal (metallic-like) gate electrode, and the down scaling of the transistor source-drain junctions calling for shallow implants and low thermal budgets. It is clear that reducing the thermal budget by performing rapid thermal annealing (RTA) is the most effective

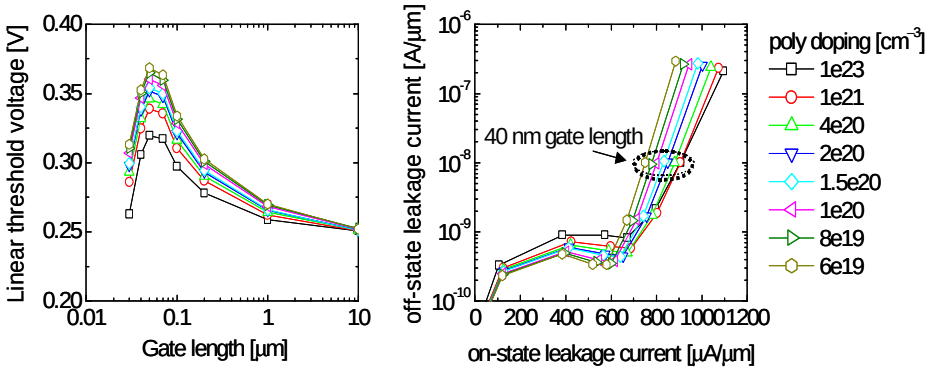
way for creating a deep-sub-micron transistor with shallow junctions ([1], [2], [3] and [4]). However, the low diffusion of the dopants results in difficulties in driving the dopants to the gate oxide interface which, in turn, results in poor gate activation.

6.3.2 Impact of polysilicon activation on NMOS transistors behavior

The impact of the activation of the polysilicon on the inversion capacitance, gate leakage, threshold voltage and drive current of NMOS devices has been investigated using device simulations [6].

In the simulations, the gate length was varied from 30 nm up to 10 µm and the doping level of the polysilicon gate was varied from 10^{23} (quasi-metallic) down to $6x10^{19}$ cm^{-3}. The polysilicon gate morphology and height (100 nm) were kept constant. The long channel threshold voltage was fixed to 250 mV by the channel doping, and the off-state current for 40 nm L_G devices was targeted at 10 nA/µm (by optimizing the halos concentration). Quantum-mechanical effects in the inversion channel were explicitly taken into account (Local Density Approximation). The physical dielectric thickness of the gate oxide was 1.2 nm.

Figure 6.3(a) and (b) show the impact of polysilicon doping on threshold voltage roll-off and on the drive current of the transistors, respectively. A higher V_T roll-up is observed in Figure 6.3(a). Indeed, in order to maintain the same I_{OFF} for all the variants, the halos concentration was changed. The SCE control is improved by increasing the polysilicon doping, where a lower halo concentration was used to maintain the same I_{OFF}. The slope of the I_{ON}-I_{OFF} curve increases with decreasing polysilicon doping level, as shown in Figure 6.3(b). The crossing of the 10^{23} and the 10^{21} curves at high current levels is a side effect of maintaining the long-channel V_T at 250 mV. The impact of polysilicon doping on transistor drive is severe: reducing the active polysilicon doping level from 10^{21} down to 10^{20} cm^{-3} decreases the on-state current by 12 %. A further reduction to $6x10^{19}$ cm^{-3} decreases the current by another 10 %.



(a) (b)

Figure 6.3: *Simulated (a) linear threshold voltage versus gate length and (b) drive current in the on-set (I_{ON}, V_{DS}=1 V) and off-set (I_{OFF}, V_{DS}=0 V) regimes for NMOSFETs with different polysilicon gate activation.*

Figure 6.4 shows the electrical oxide thickness (T_{ox}^{elec} as defined in (6.1)) dependence on the active dopant concentration in the polysilicon gate (assuming a uniform doping profile):

$$T_{ox}^{elec} = T_{ox}^{phys} + T_{ox}^{qm} + T_{ox}^{dp} \tag{6.1}$$

where T_{ox}^{phys} is the physical oxide thickness of the dielectric layer, T_{ox}^{qm} is the additional thickness due to quantum confinement and T_{ox}^{dp} is the additional thickness due to polysilicon depletion.

The change in T_{ox}^{elec} was determined from the inversion capacitance at $V_{GS}=V_T+1$ V on simulated 10 μm long NMOS transistors and compared to values obtained from a device without polysilicon depletion (*i.e.* quasi-metallic gate, 10^{23} cm^{-3} polysilicon doping).

Reducing the polysilicon doping concentration from 10^{21} down to 10^{20} cm^{-3} corresponds to 0.6 nm electrical oxide thickness increase, which corresponds to more than 30 % of the total electrical thickness required in advanced CMOS technology nodes (T_{ox}^{elec}< 2 nm)! Note however, that the relative decrease in drive current (Figure 6.3(b)) is significantly less than the relative increase in T_{ox}^{elec}. The reason for this is twofold: (i) Since carriers in the inversion layer experience higher vertical electric fields in the case without polysilicon depletion, their mobility is lower; (ii) quantum effects (i.e. inversion charge not being localized at the silicon/dielectric interface) give rise to an increase of T_{ox}^{elec}. For a T_{ox}^{phys} of ~1.2 nm, quantum effects in the channel add about 0.4 nm to the T_{ox}^{elec}, and this contribution increases with decreasing T_{ox}^{phys}.

Moreover, the increase of T_{ox}^{elec} for lower polysilicon doping levels has only limited impact on the gate current density. 0.2 nm increase of T_{ox}^{elec} only leads to 20-30 % decrease in the gate leakage current density (J_{gate}) (to be compared with 1 decade decrease in J_{gate} for 0.2 nm increase in T_{ox}^{phys}). Since the gate current density is determined by direct tunneling, it depends strongly on the physical oxide thickness and only moderately on the potential drop over the dielectric.

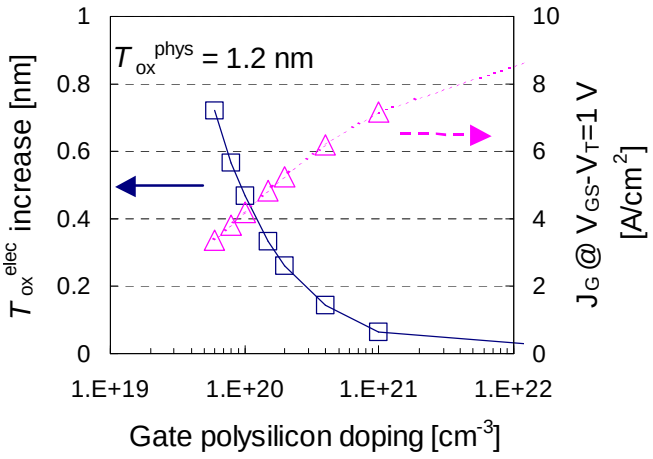


Figure 6.4: T_{ox}^{elec} and J_G dependence on polysilicon doping for a long channel (L_G=10 µm) NMOST featuring a 1.2 nm T_{ox}^{phys} gate dielectric.

As a consequence of the strong dependence of the gate active doping concentration on the electrical thickness of the gate dielectric, it is of first importance to maximize the activation of the gate electrode.

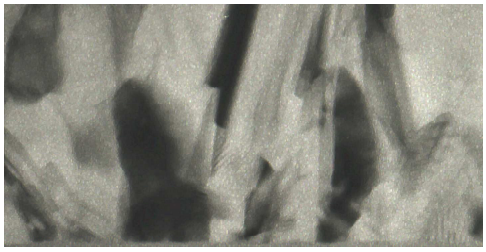
6.3.3 Influence of the gate electrode morphology on dopants activation

It has been reported that the gate polysilicon morphology influences the dopants diffusion [7], [8]. In this section, variations in gate morphology and in post-implant RTA conditions are studied in detail with the aim of finding a process window with high gate activation: the beneficial effect of a spike anneal (0 sec) was confirmed ([9], [10] and [11]).

The activation behavior of polysilicon layers was studied by depositing a 150 nm thick silicon film on top of a thick 3.5 nm oxide layer by low pressure chemical vapor deposition (LPCVD) using SiH_4 as the reactant gas. The grain size and structure of this film was controlled by varying the reactor temperature during growth: polysilicon films were deposited at 610 °C and amorphous silicon films at 550 °C. The amorphous silicon layers were then recrystallized through annealing at 850 °C during 15 minutes in nitrogen ambient. After gate patterning, the source/drain extensions were implanted and activated, and the spacers were formed. The highly doped source drain regions (HDD) and thus the final doping of the gate were made by implanting 40 keV arsenic (As) or 3 keV B followed by a 20 s RTA at temperatures ranging from 960 to 1050 °C in nitrogen. No silicidation was performed

in order to minimize a possible influence of dopant deactivation. C-V measurements were performed on $300 \times 300\ \mu m^2$ capacitor structures.
Transmission electron microscopy (TEM) cross-section views of polysilicon and amorphous/recrystallized films are shown in Figure 6.5. A very different grain structure is observed between the two gate materials: the polysilicon material exhibits a columnar structure with grains of 30-100 nm in diameter, while the amorphous/recrystallized material shows bulky grains of 100-300 nm in diameter. This structure difference will influence the diffusion of the dopants, since dopants tend to diffuse rapidly along grain boundaries while longer time is required to diffuse the dopants into the polysilicon grains [12].

(a)



(b)

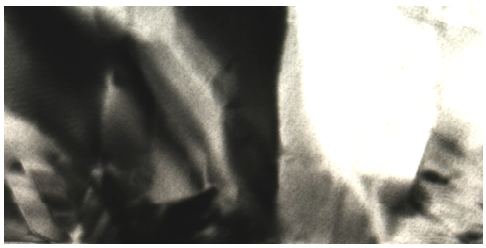


Figure 6.5: *Transmission electron micrograph of (a) fine-grained polysilicon deposited at 610 °C and (b) amorphous silicon deposited at 550 °C and recrystallized at 850 °C for 15 minutes.*

The active dopant concentration at the interface gate electrode/gate oxide of PMOS and NMOS transistors was evaluated by calculating, from C-V measurements, the gate depletion, using the following convenient equation (that holds well for thick dielectrics when quantum mechanical effects are insignificant):

$$D_{gate} = \frac{C_{acc} - C_{inv}}{C_{acc}} \times 100\% \qquad (6.2)$$

where D_{gate} is the level of gate depletion, C_{acc} and C_{inv} are the capacitance measured under accumulation and inversion bias settings at ± 2 V, respectively. This expression of the gate depletion can be rewritten in terms of the capacitance across the oxide (C_{ox}) and the capacitance across the depleted part of the polysilicon gate ($C_{depletion}$). With the approximations $C_{acc} = C_{ox}$ and $C_{inv}^{-1} = C_{ox}^{-1} + C_{depletion}^{-1}$, equation 1 becomes:

$$D_{gate} = \frac{C_{ox}}{C_{ox} + C_{depletion}} \times 100\% \quad (6.3)$$

The gate depletion of the two gate materials as a function of the activation temperature is shown in Figure 6.6 for both N- and PMOS devices.

The behavior of the amorphous/recrystallized silicon electrode is identical for the NMOS and PMOS with a minimum gate depletion of 5 % obtained at 1030 ^{o}C. At low thermal budget (T<1020 ^{o}C), the gate active impurity concentration at the gate/gate oxide interface is much higher for the fine-grained polysilicon. This can be attributed to the small size of the grains and therefore to a large amount of grain boundaries where the diffusion process mainly occurs which leads to a higher active impurity concentration at the gate/gate oxide interface ([12], [13]).

For NMOS transistors, increasing the temperature (T>1020 ^{o}C) results in worse gate depletion for the polysilicon gate structure as well as for the amorphous/recrystallized silicon one. This result is consistent with the behavior of the arsenic dopants, activated at high temperature, which are segregating at the Si-boundaries and/or evaporating from the gate to the ambient (although this latter should not affect the doping concentration at the gate electrode/gate dielectric interface), [14], [15], [16] and [17].

For PMOS transistors having a polysilicon gate electrode, a rather constant gate depletion is measured independently of the anneal temperature. Moreover, higher polysilicon depletion, that is to say lower active impurity concentration at the gate/gate oxide interface, is observed for the p-type polysilicon gate as compared to the n-type one. This could be attributed to the fact that while As dopants segregate at the polysilicon gate/gate dielectric interface, B atoms are diffusing through the gate dielectric. The B concentration at the interface gate electrode/gate dielectric is therefore lower than the implanted peak concentration, yielding higher gate depletion.

It can be concluded that the use of fine-grained polysilicon results in a decrease of the polysilicon depletion. We have shown that maximizing the gate activation while reducing the thermal budget of the RTA is possible in using fine-grained polysilicon gates for both N- and PMOS transistors.

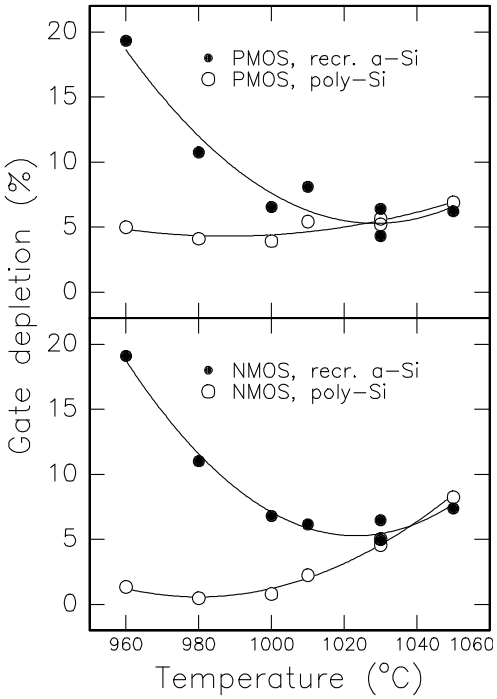


Figure 6.6: *Gate depletion in NMOS and PMOS transistors as a function of the source/drain activation temperature (20 s anneal). The gate dielectric is a 3.5 nm pure oxide film. Two types of gate materials are compared: polysilicon (poly-Si) and amorphous/recrystallized silicon (recr. a-Si).*

6.3.4 Optimization of dopants activation in the polysilicon gate

As presented in the previous section, fine-grained polysilicon, as the gate electrode material, yields very good dopants activation. Yet, achieving the very high dopants activation required for sub-100 nm CMOS technologies remains a real challenge [22]. Increasing the polysilicon gate doping at or near the gate/gate dielectric interface will result in a measurable decrease in the polysilicon depletion layer thickness and therefore in an increase in device performance. To increase the doping in the gate it is necessary to place a large concentration of active dopants near the polysilicon gate/gate oxide interface, and to maintain this level of activation. Since junction depths are scaling faster than the polysilicon gate height, in many cases it is not desirable to dope the polysilicon gate with the same implants and thermal cycles used to dope the deep source/drain (junction anneal). This is particularly true for NMOS devices in which As is used for the deep source/drain implants. Taking these remarks into consideration, solutions to improve the polysilicon activation are proposed in this section.

6.3.4.1 Pre-doping the polysilicon gate

Special doping and activation anneal processing steps (so called pre-doping process) were added prior to the source/drain extensions formation to fully activate the polysilicon gate, as first proposed by M. Rodder et al. [23] (see also [24]). A comparison of the C-V characteristics measured on NMOS devices with or without a pre-doping step is presented in Figure 6.7(a). The fine grained polysilicon gate height is 100 nm, the pre-doping is a phosphorous (P) implant (20 keV, 3×10^{15} cm^{-2}) annealed at 950 ^{o}C for 30 sec. P is preferred over As as it is a smaller atom with a high diffusivity enabling the peak concentration to be closed to the interface polysilicon/gate dielectric. It can be observed that in the inversion regime, the capacitance is strongly attenuated if no pre-doping is performed. In this case, the pre-doped polysilicon gate yields 12 % higher polysilicon activation. In Figure 6.7(b), the linear and saturated V_T for NMOS transistors with fine-grained polysilicon with or without pre-doping at various gate lengths is presented. It can be observed that for short channel devices, the V_T becomes dependent on the gate length: this is the so-called V_T roll-off (i.e. the difference between long channel V_T and short channel V_T). The V_T roll-off starts at a longer channel length for the non pre-doped polysilicon gate compared to the pre-doped one. This is in good agreement with the fact that, because of reduced polysilicon depletion, better control of the gate on the channel is obtained. In other words, better SCE control is obtained when pre-doping the n+ polysilicon gate.

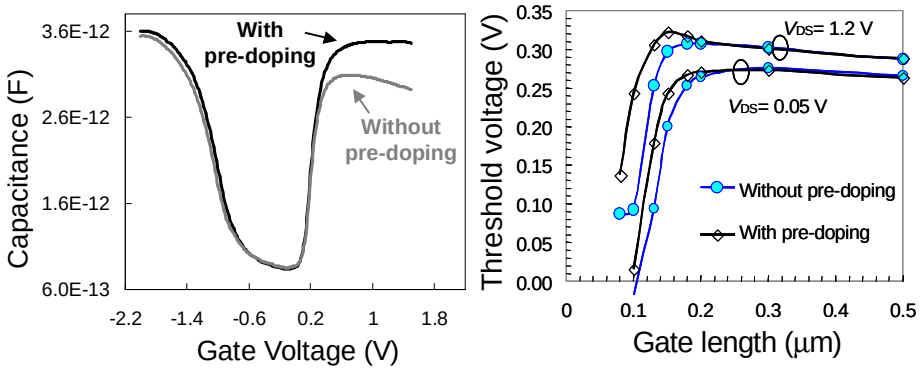


Figure 6.7: *Comparison of NMOS devices with or without a gate pre-doping step. The gate dielectric is a plasma nitrided oxide with an EOT of 1.5 nm. (a) C-V characteristics for NMOS capacitor and (b) Linear and saturated V_T for NMOS devices at various gate lengths.*

We have shown that pre-doping the n+ polysilicon gate yields better transistor performance. The next logic step is to investigate the impact of pre-doping the polysilicon gate of PMOS transistors. The junctions of a PMOS transistor are usually

formed using Boron (B) dopants. Contrary to As, B atoms can diffuse quickly through the polysilicon grain boundaries and can also diffuse through the gate dielectric into the substrate. This phenomenon is called B penetration yielding a change in the doping of the channel and therefore a shift in V_T ([25] and [26]). It was reported in the literature that nitrogen in the gate dielectric is an effective diffusion barrier for B atoms to diffuse into the channel ([27], [28] and [29]). The linear V_T of PMOS devices with a lightly doped nitrided (using a furnace nitridation process) and a highly doped nitrided (using a plasma nitridation process, namely DPN) gate oxide have been compared at various activation anneal temperatures. The p+ polysilicon gate was not pre-doped. The EOT of both gate oxides is 1.5 nm EOT. It can be observed in Figure 6.8(a) that the linear V_T of PMOS devices with a plasma nitrided gate oxide is rather constant over the whole anneal temperature range while the V_T of PMOS devices with a furnace nitrided gate oxide is increasing dramatically with the temperature. This large increase in V_T is the signature of B penetration in the channel through the thin furnace nitrided oxide. Having more N in the oxide yields a significant decrease in B penetration. This result underlines another benefit of using plasma nitrided oxide instead of pure oxide or lightly nitrided oxide as the gate dielectric. The difference in V_T between the PMOS devices with the two gate dielectrics can be attributed to the difference in N concentration in the gate oxide. As already explained in the previous chapter, the incorporation of large amount of N yields the creation of positive fixed charges that will increase the absolute value of V_T of PMOS transistors. Nevertheless, the value of V_T can be tuned by adjusting the dopants in the channel.

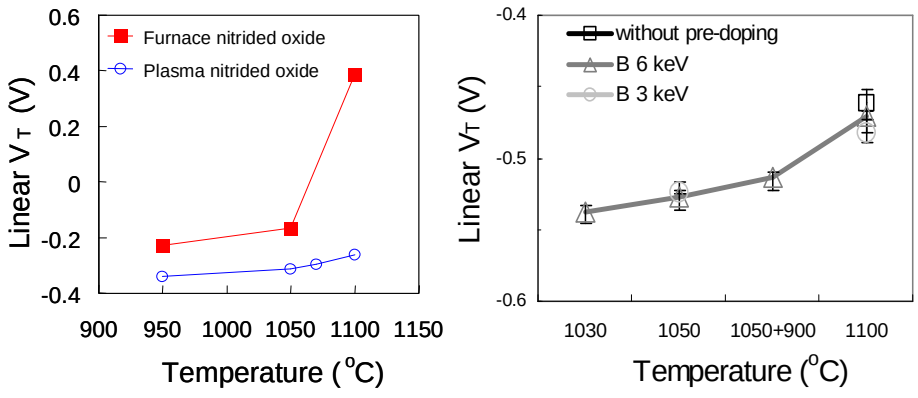
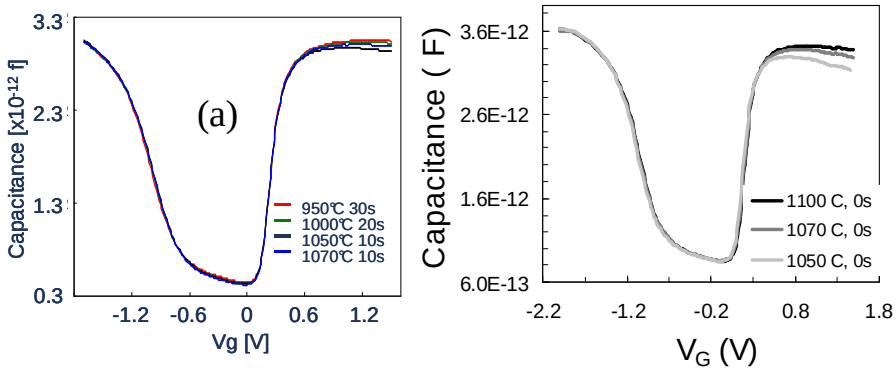


Figure 6.8: *(a) Linear V_T as a function of junctions anneal temperature for long channel PMOS devices with a lightly nitrided (furnace nitrided) or heavily nitrided (plasma nitrided) gate oxide. The EOT of the two gate dielectrics is 1.5 nm. (b) Linear V_T for long channel PMOS devices with or without pre-doped gate. For the pre-doping gate devices, the B dose for the implant is fixed at 1×10^{15} at.cm^{-2}. The*

gate dielectric is a 1.5 nm EOT plasma nitrided oxide.

A pre-doping implant was then performed on PMOS devices with fined grained polysilicon and a plasma nitrided oxide. Two pre-doping implant conditions were studied: B 1×10^{15} at.cm^{-2} at 3 or 6 keV. A comparison of the long channel linear V_T of PMOS devices with these pre-doping implant conditions is shown in Figure 6.8(b) for various anneal temperatures. The linear V_T of PMOS devices without pre-doping are also plotted as a reference. Similar linear V_T is measured for all p+ polysilicon gates. This indicate that pre-doped p+ polysilicon gate using a 3 or a 6 keV B implant can be employed without enhancing B penetration phenomenon. Similarly to the impact of pre-doping the n+ polysilicon gate, better SCE control has been observed on PMOS devices with a pre-doped polysilicon gate.

Finally, the anneal following the pre-doping implant has been investigated. Figure 6.9(a) shows the C-V curves measured on n-type polysilicon transistors pre-doped with 20 keV P and annealed at temperatures ranging from 950 °C to 1070 °C. It can be observed that, for a given gate dielectric (plasma nitrided oxide, EOT=1.5 nm), polysilicon gate height (150 nm) and pre-dope implant, varying the pre-dope anneal has a small impact on the C-V characteristics. The polysilicon depletion (evaluated from the difference of the accumulation and inversion regimes, as already presented in Equation 6.3) is therefore rather similar whatever the pre-dope anneal. Nevertheless, the data demonstrates that the best condition for this anneal is 950 °C, 30 s and indicates that in this case the pre-dope anneal distributes the dopants in the gate, while the final junction anneal is responsible for dopants activation.



(a) (b)

Figure 6.9: *C-V characteristics for NMOS capacitors with a 1.5 nm EOT plasma nitrided oxide and a pre-doped polysilicon gate with (a) various pre-dope annealing conditions (junctions anneal is 1050 °C, spike) and (b) various HDD spike anneal temperatures (pre-dope anneal*

is 950 °C, 30 sec).

6.3.4.2 Optimization of the thermal budget

The HDD anneal temperature has been here investigated with the view to maximize dopant activation in the gate. We will see later in this chapter that there is a trade-off between fully activating the gate and obtaining ultra-shallow junctions. In Figure 6.9(b), the C-V characteristics of NMOS transistors having received different activation anneal (junctions anneal) have been measured. Spike anneal (0 sec anneal) were performed at 1050, 1070 or 1100 °C. The n+ polysilicon gate were all pre-doped and annealed at 950 °C during 30 sec. A measurable decrease in the polysilicon depletion is observed when the spike anneal temperature is increased from 1050 °C to 1100 °C, indicating that more active dopants exist near the polysilicon/oxide interface. Therefore, in order to maximize the polysilicon gate activation, high temperature HDD anneal is required.

6.3.4.3 Optimization of the polysilicon gate height

An easy way to get a large amount of dopants at the interface gate/gate dielectric is to reduce the height of the polysilicon gate. When the height of the polysilicon is decreased, the level of polysilicon depletion is also decreased, as shown in Figure 6.10. With a reduced gate height, dopants are implanted closer to the polysilicon/gate oxide interface, for both the pre-doped implant and the HDD implants.

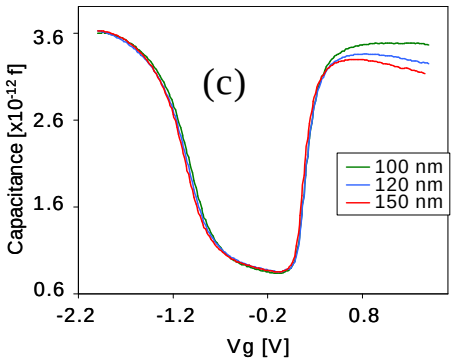


Figure 6.10: *C-V characteristics for n+ polysilicon gate having various heights.*

6.3.4.4 Minimization of dopants deactivation

It has been shown in this chapter that by carefully optimizing polysilicon gates and activation anneals, gate activation can be maximized. It was explained that the polysilicon gate exhibits the best activation mainly due to its fine-grained structure, which facilitates dopant diffusion. Unfortunately, this "easy" diffusion can turn out to

be a major drawback when considering the extra annealing steps (silicidation and metallization steps) in a full CMOS process. Further thermal processing can indeed make the dopants segregate at the grain boundaries, evaporate or diffuse through the oxide (Boron penetration phenomenon) [18]. This loss of activation may induce a severe increase in the level of gate depletion and thus a degradation of transistor performance.

Gate deactivation kinetics

The aim of this section is to study and compare the deactivation kinetics of the gate, both p- and n-type, using polysilicon or amorphous/recrystallized silicon. After a post-implant anneal at 1030 ^{o}C during 20 s, the samples were annealed at different times and temperatures. The average Hall carrier concentration for the two gate structures was measured as a function of the annealing time at 700, 750 and 800 ^{o}C. In Figure 6.11, the exponential decrease of the carrier concentration is plotted as a function of annealing time for both structures and dopants types. Note that these values of carrier concentrations are inaccurate due to the unknown value of the Hall factor [19]. The observed decrease of the carrier concentrations tends to a value, which is related to the solid solubility of the dopants in the gate at the annealing temperature. The higher the anneal temperature, the faster this saturation level is reached as a result of a faster diffusion process. To describe this deactivation mechanism in a more quantitative fashion, the carrier concentration has been fitted using:

$$C(t) = [(C_{initial} - C_{final}) \times \exp(-\frac{t}{\tau}) + C_{final}] \qquad (6.4)$$

where C_{initial} and C_{final} are the active carrier concentrations calculated respectively after the post-implant anneal and at the saturation level. The time constant (τ), which describes the kinetics of the gate deactivation [20] can be calculated from (6.4) and is presented in the Arrhenius plot in Figure 6.12.

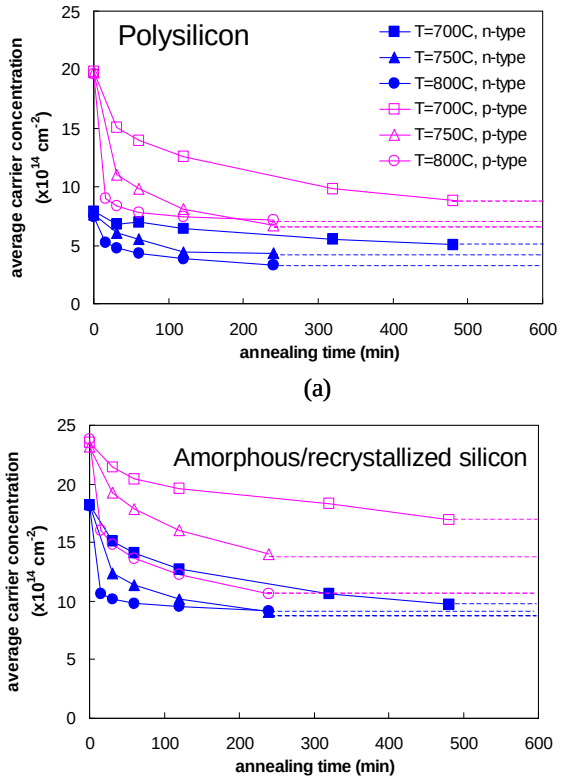


Figure 6.11: *Hall carrier concentration as a function of annealing time and temperature for (a) fine-grained polysilicon and (b) amorphous/recrystallized silicon. The films are B or As doped. Annealing temperatures (in °C) are indicated in the legend common to the stwo figures.*

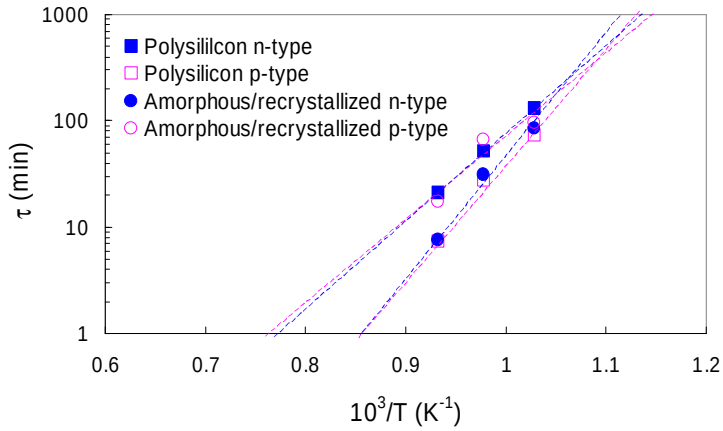


Figure 6.12: *Arrhenius plot of the dopant deactivation time constant for n- and p-type dopants and different gate morphologies.*

It can be observed that τ is strongly dependent on temperature. The activation energy for the polysilicon and the amorphous/recrystallized silicon gate, both p- and n-type are reported in Table 6.1. For the p-type doped gates, the polysilicon structure exhibits higher activation energy and a faster deactivation than the amorphous/recrystallized silicon gate material. The large amount of grain boundaries of the polysilicon structure gives rise to a faster deactivation: the boron atoms can diffuse very easily and be trapped at grain boundaries, causing deactivation. For the n-type gates, however, the deactivation is faster for the amorphous/recrystallized silicon material in spite of the bulky grain structure. We did not find so far a physical explanation to this result.

s	Polysilicon p-type	Polysilicon n-type	Am/recryst Si p-type	Am/recryst Si n-type
Activation energies (eV)	2.07 ± 0.24	1.67 ± 0.03	1.52 ± 0.54	2.17 ± 0.26

Table 6.1: *Activation energies calculated for polysilicon and amorphous/recrystallized silicon films p- and n-type doped.*

Based on these results, the gate deactivation was calculated. Figure 6.13 shows the gate activation loss calculated after each thermal step of a CMOS process flow.

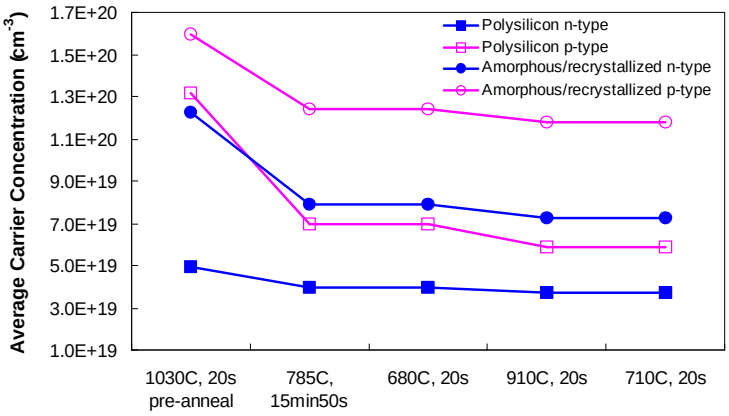


Figure 6.13: *Calculated activation loss for a fine-grained polysilicon and an amorphous/recrystallized silicon structure after various anneal steps of a CMOS process flow. Both n- and p-type doped gates are compared.*

The major contribution of the activation loss is the nitride deposition (785 ^{o}C, 15 min 50 s) used as a silicide protection layer prior to silicidation. This deposition could be done by Plasma Enhanced Chemical Vapor Deposition (PECVD) instead of LPCVD at a lower temperature (400 ^{o}C), to avoid gate deactivation related to this step. In this experiment, the silicidation process itself consists of the formation of titanium silicide ($TiSi_2$) as a contact on the polysilicon gate and active areas. Because the titanium is a very reactive atom, it can form inactive compounds with the dopants and thus increase the gate deactivation at the interface gate/silicide. The thermal treatment used for this step can also be harmful for the gate activation. The maximum temperature for this step is high (910 ^{o}C). Calculations of the deactivation percentage due to the silicide step thermal budget have revealed an activation loss at the maximum of 5 to 20 % depending on dopants and gate material. These average values are not necessarily representative of the amount of active dopants at the gate/gate oxide interface and it is thus difficult to evaluate the level of gate depletion. Transistors with a physical gate oxide thickness of 3.2 nm (electrical thickness: 4 nm) have been measured and exhibit an increase in gate depletion due to silicidation from 2 to 4 % for NMOS transistors, and from 4 to 6 % for PMOS transistors. In this experiment, these gate depletion levels are still acceptable for good transistor performance but the challenge is to get such a level for advanced CMOS generations (≤100 nm), employing thinner gate oxides. One obvious route to minimizing dopants deactivation is the use of lower thermal budget for post annealing steps (after HDD RTA step). The impact of reducing the thermal budget of the silicidation step on the polysilicon depletion level will be shown in the next section of this chapter.

Silicidation optimization for gate deactivation reduction

Because of its high thermal budget, the silicidation step has been identified in the previous paragraph as one of the most harmful step for deactivation of dopants. In that experiment, $TiSi_2$ silicide was used having a maximum temperature of 910 ^oC. In this paragraph, the silicidation step will be optimized with a view to decrease the polysilicon depletion of both N- and PMOS transistors.

Nickel silicide (NiSi) is an attractive material because it consumes less silicon than $CoSi_2$ or $TiSi_2$ and has lower processing temperatures (maximum temperature is 450 ^oC for NiSi, 850 ^oC for $CoSi_2$ and 910 ^oC for $TiSi_2$) while having a similar resistivity [30].

A comparison of the polysilicon gate depletion was made for N- and PMOS transistors formed with a high temperature silicidation (in this case $CoSi_2$) and a low temperature silicidation process (namely NiSi). An increase of the inversion capacitance is observed for the PMOS device having the NiSi process step, as presented in Figure 6.14. This decrease in polysilicon depletion results from the reduction in polysilicon dopant deactivation that occurs during the low temperature Ni process. After high temperature junction activation (in our case: spike anneal at 1050 °C), the dopants in the polysilicon gate (as well as in the bulk silicon) are susceptible to deactivation during subsequent thermal processing (see section 6.3.3.2). That is, thermal treatments below the high temperature activation step, such as the silicidation anneal, will drive the system (dopants in silicon) towards the equilibrium active doping concentration (solubility) at that temperature. The 850 °C thermal treatment of the Co silicide is considered a significant enough anneal to push the dopant system towards the new thermal equilibrium, i.e. to a lower doping level than that at 1050 °C. Although the active dopant equilibrium level (solubility) at 450 °C is significantly lower than at 850 °C, this thermal treatment during Ni silicidation is not long enough to cause significant movement of the systems towards thermal equilibrium. Therefore, more dopant deactivation during the 850 °C Co silicidation process is expected resulting in higher polysilicon depletion level. NMOS devices do not show appreciable difference in polysilicon activation levels due to their slow diffusivity and higher solubility level (Figure 6.14).

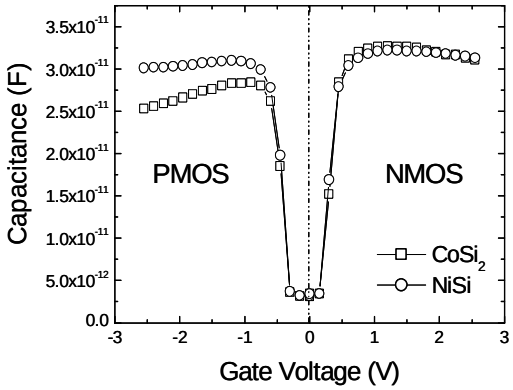


Figure 6.14: *Comparison of C-V characteristics (from depletion to inversion) for both P- and NMOS devices with a $CoSi_2$ (850 °C) or NiSi (450 °C) silicidation step.*

As a result of the change in polysilicon activation level, PMOS transistors with NiSi exhibit a higher drive current than devices with $CoSi_2$. An increase of about 30 % is observed in the peak of the transconductance (Figure 6.15) at a given bias, V_G-V_T. Very small improvement in NMOS device performance (less than 5 %) was observed probably attributed to slightly lower polysilicon depletion.

As a conclusion to this section, it has been shown that fine-grained polysilicon gate with pre-dope implant processing step yields high gate activation. High RTA thermal budget is required to maximize the activation of the dopants at the polysilicon gate/gate dielectric interface. Reducing the thermal budget of the silicidation step can minimize dopants deactivation. The replacement of $TiSi_2$ and $CoSi_2$ by NiSi results in minimum polysilicon depletion.

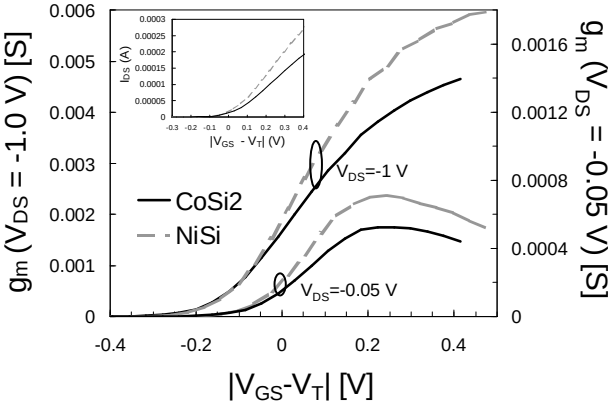


Figure 6.15: *Transconductance versus gate overdrive for PMOS devices (W=10 μm, L_G=66 nm) having a plasma nitrided oxide of 1.4 nm EOT and processed with a $CoSi_2$ (850 °C) or NiSi (450 °C) silicidation step. The drive current of these PMOS transistors is plotted in the inset of the figure.*

6.4 Compatibility of optimized polysilicon gate stack with advanced ultra-shallow junctions

We have seen in section 6.3 that minimum polysilicon depletion is observed when using high temperature spike anneal RTA. However, the formation of ultra-shallow junctions require minimum thermal budget for this activation anneal ([35], [36]). There is therefore a trade-off between getting a fully activated polysilicon gate and forming ultra-shallow junctions. In this section, the compatibility of optimized gate stack (including ultra-thin plasma nitrided oxide and highly activated polysilicon gate) with the formation of ultra-shallow junctions (USJ) will be investigated. The impact of new techniques to form USJ on the gate stack will be discussed.

6.4.1 Introduction to ultra-shallow junctions

6.4.1.1 Motivation

When the device is scaled, the source/drain junctions have to be scaled to avoid punch through. Punch through occurs when the source and drain space charge regions overlap and results in uncontrolled SCE. Reducing the source/drain junction depth (X_j) and lateral profile abruptness will improve device short channel characteristics by reducing the amount of channel charge controlled by the drain [37]. In order to get good short channel transistor behavior, a shallow lowly doped junction close to the channel (LDD) is formed prior to a deeper highly doped junction (HDD), located further away. The USJ (LDD junctions) implants are formed just after the halo implantation step, using an ultra-low energy implant.

6.4.1.2 Scaled junctions requirements

There are three important parameters to optimize when forming USJ: X_j (measured at a concentration of dopants of 10^{18} cm^{-3}), the sheet resistance (R_{sheet}) and the abruptness of the junctions.
USJ yield an increase of R_{sheet} that adds to the total external resistance degrading the saturated drain current ([38] and [39]). There is therefore a trade-off between getting shallow source/drain junctions for control of SCE and having lowly resistive junctions to get high device performance.

The abruptness of the USJ has also an impact on the device performance. The on-state drive current (I_{ON}) of NMOS transistors having USJ of different abruptness has been simulated for two gate-to-drain overlap [6]. The lateral abruptness was calculated as the slope of the dopants profile. The off-state drain current was kept constant. As can be observed in Figure 6.16, more abrupt junctions result in higher drive current, at fixed I_{OFF}. This can be explained by the fact that abrupt shallow junctions will reduce the accumulation resistance (R_{acc}). This will result in an increase of I_{ON}. The main effect of making junctions more abrupt is to decrease the effective channel length. This will enable better SCE control and avoid I_{ON} degradation. As a consequence, the gate-to-drain overlap can be decreased while still yielding similar drive current with very good SCE control (Figure 6.16).

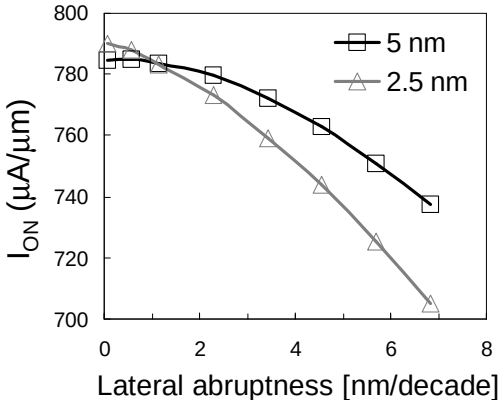


Figure 6.16: *Simulated I_{ON} as a function of the lateral abruptness of the LDD junctions. The I_{OFF} is kept constant. Two junction to gate overlap lengths have been simulated (5 and 2.5 nm at each side of the channel).*

Therefore an ideal "box-like" profile for the junctions, as required for further MOS downscaling ([22]), will enable a reduction of the gate-to-drain overlap. By decreasing the overlap, the source-to-drain capacitance can be reduced. Another consequence of the reduced gate-to-drain overlap is that the off-state gate current (I_{G_OFF}), as defined in chapter 2, will be decreased. The I_{G_OFF} has been calculated at various EOT (taking the gate leakage current values obtained for plasma nitrided oxides) and for various gate-to-drain overlap lengths, as presented in Figure 6.17. A linear decrease of I_{G_OFF} with the gate-to-drain overlap is observed. This small decrease is of course less significant than the exponential decrease of I_{G_OFF} when scaling the dielectric thickness.

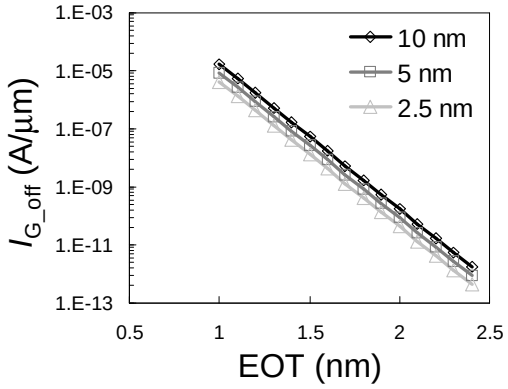


Figure 6.17: *off-state gate leakage current (J_G times the gate-to-drain overlap) as a function of the EOT for various gate-to-drain overlap lengths.*

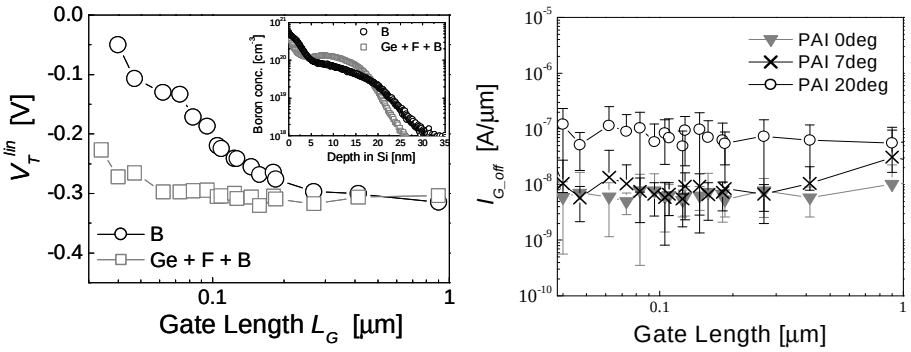
While USJ for NMOS transistors have been reported that satisfy the X_j/R_{sheet} requirements ([22]) for sub 30 nm gate length transistors ([40]), forming USJ for PMOS devices remain a real challenge. Conventional ultra-low energy implant with a spike RTA do not meet the X_j - R_{sheet} and abruptness specification ([22]) for short channel PMOS transistor (L_G<45 nm), as reported in the literature [44].
Therefore, alternative techniques to conventional implantation and rapid thermal annealing have to be found to form p-type USJ.

6.4.2 Alternative techniques to form ultra-shallow junctions and compatibility with the gate stack

6.4.2.1 Pre-amorphized junctions with spike annealing

Pre-amorphized junctions are currently studied to form p-type ultra-shallow junctions. Extensive investigations have been carried out on pre-amorphization implants (PAI); co-implantations in combination with low-thermal budget (high or low temperature) anneal to provide highly active and abrupt USJ ([41]- [51]). It has been shown that by use of deep Germanium (Ge) pre-amorphization, properly tuned Fluorine (F) co-implantation and fast ramp-up (RU) and ramp-down (RD) high-temperature anneals, Boron (B) USJ can be obtained ([42]-[46], [48]-[51]). F is used to reduced B Transient Enhanced Diffusion (TED). A comparison of the linear V_T of PMOS devices having conventional B junctions or PAI B junctions with F co-implantation is presented in Figure 6.18(a). An almost flat V_T is obtained with the PAI B junctions indicating that better SCE control is obtained with such junctions. The B dopant profile obtained for the two junctions is shown in the inset of Figure 6.18(a). This is in good agreement with the measured B profile using SIMS analysis. As can be observed in the inset of Figure 6.18(a), the PAI F co-implanted B junction

profile is more abrupt than the conventional B junction, resulting in better short channel characteristics.



(a) (b)

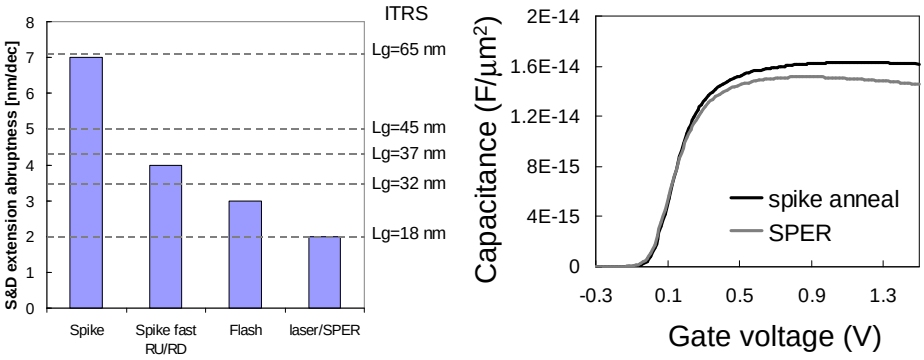
Figure 6.18: *(a) Comparison of V_{T_lin} for PMOS devices with optimized pocket implants and ultra-shallow source/drain extension implants using B or using PAI F co-implantation. (b) I_{G_off} for PMOS transistors having PAI F co-implanted B junctions at various Ge implantation tilts. The RTA is a 1070 °C spike anneal.*

To investigate the impact of the pre-amorphization on the gate stack, the I_{G_OFF} has been compared for PMOS devices with PAI F co-implanted B junctions activated using a spike anneal at 1070 °C. Several tilts for the Ge implant have been investigated, resulting in different drain-to-gate overlap lengths. In Figure 6.18(b), a comparison of I_{G_OFF} for three different tilts of the Ge PAI is presented. Similar I_{G_OFF} has been measured for low tilt PAI. However, a significant increase in I_{G_OFF} is measured when the PAI is done a high tilt of 20°. This increase can be partly explained by the gate-to-drain overlap increase but also might be related to the damage induced to the gate dielectric by the heavy PAI dose used to create the junctions (5×10^{14} at.cm^{-2}). It seems that the damage induced by the tilted implants cannot be healed even by a high-temperature junction activation anneal step.
As a conclusion, while shallow abrupt B junctions can be manufactured by Ge pre-amorphization and F co-implantation, the gate dielectric might be damaged if the implantation tilt is not well tuned [51].

6.4.2.2 Solid Phase Epitaxial Regrowth junctions

Other alternatives to conventional implantation and activation anneal is the use of low temperature Solid Phase Epitaxial Regrowth (SPER, [52], [53]), flash anneals with and without co-implantations (F, Ge, C, N, …) and laser annealing (LTA, full melt [54]). They are currently widely investigated to overcome some of the limitations of standard RTA processes. The main advantages of these techniques are high dopant activation levels (above solid solubility) with minimal diffusion. The lateral abruptness of junctions formed with these various annealing techniques has been measured on SIMS B dopant profiles. SPER together with laser annealing

techniques yield to the most abrupt junctions, as shown in Figure 6.19(a). The impact of SPER junctions on the polysilicon gate stack has been here studied. The thermal budget of the SPER process was 650 ^{o}C for 1 minute. It was observed that the gate leakage current was similar for N- and PMOS devices with conventional or SPER junctions. However, a decrease in the inversion capacitance has been observed for NMOS devices with SPER junctions, resulting in a 0.2 nm increase in CET, as can be observed in Figure 6.19(b). A possible explanation for this increase in polysilicon depletion could be that during the pre-amorphization phase, part of the polysilicon gate is also amorphized. This pre-amorphization will create End-Of-Range (EOR) defects located at the bottom of the gate that will not be healed with the regrowth anneal of 650 ^{o}C for 1 minute. These defects will deactivate dopants in the gate and particularly dopants close to the interface with the gate dielectric. As a result, the polysilicon depletion will be increased. Another possible explanation might be the deactivation of dopants in the gate during the thermal regrowth of the junctions.



(a) (b)

Figure 6.19: *(a) p-type junction abruptness for various annealing techniques. Comparison with ITRS lateral abruptness requirements for given gate length [22]. (b) C-V characteristics for NMOS devices with conventional spike annealing at 1050 ^{o}C or SPER junctions with a regrowth anneal of 650 ^{o}C for1 minute.*

6.5 Conclusions

The impact of scaling the gate dielectric thickness on short channel transistor performance has been shown. Ultra-thin plasma nitrided oxides have been successfully integrated in sub-50 nm devices yielding good transistor performance with limited leakage current. A study of the dopant activation and deactivation in the gate electrode has resulted in the choice of fine-grained polysilicon structure. Gate pre-doping implantation, optimized RTA and low temperature silicidation process yield highly activated polysilicon gate with low polysilicon depletion. Finally, the compatibility of the optimized gate stack with advanced ultra-shallow junction

formation techniques has been investigated. It was shown that while pre-amorphization and SPER junctions result in shallow abrupt junctions, they could have a significant impact on the gate dielectric and polysilicon activation.

The solid solubility limit of dopants in polysilicon and the dopant deactivation resulting from mid-temperature processing limit the level of activation achievable. Therefore, in order to achieve the significant performance increase for short channel devices, it is necessary to change the material properties of the gate. Metallic gates are currently under investigation, as they will eliminate polysilicon depletion effects, solve the problem of B penetration and would be more compatible with the formation of abrupt ultra-shallow junctions.
The use of metals with appropriate work functions for NMOS and PMOS devices (4.1-4.2 and 5.2-5.3 eV, respectively) would lead to transistors with symmetrical and tailored V_T [32], [33]. V_T fluctuations due to statistical doping variations in the gate will be strongly reduced. Other significant advantages to use metal gate electrode are the reduction of the gate resistance (needed for high speed applications) and the elimination of the Coulomb carrier scattering from doping variations and polysilicon grains in the gate electrode. Finally, the integration of new high-*K* dielectrics could be facilitated in using a metal electrode, although this is not necessarily true for all the electrode materials that could be of interest. In addition, it is currently unclear whether the "pinning" of the Fermi-level is also an issue for High-*K* in combination with metal gate. Furthermore, the thermodynamic stability of the metal gate/gate dielectric interface at processing temperatures is a major concern that needs to be addressed in addition to the more subtle issues of electrical properties, flat band voltage (ultimately threshold voltage) stability and charge trapping at the interface [34]. Furthermore, physical damage to the dielectric during sputtering or deposition of the metal films and metal diffusion through the gate dielectric are major processing issues that need to be considered.
Other challenges related to bulk Si technology are the use of gate electrodes with work functions separated by roughly the band gap of Si for NMOS and PMOS devices in order to obtain low and symmetrical threshold voltages in NMOS and PMOS devices and minimize short channel effects. Moreover, the use of two refractory metals introduces the additional complexity of etching, selectivity and suitable masking procedures to selectively deposit metals over different areas of the same wafer.

6.6 References

[1] F. Stern and W.E. Howard, *Self-Consistent Results for n-Type Si Inversion Layers*, Phys. Rev. Vol. 163, pp. 816-835, 1967.

[2] F. Stern, *Self-Consistent Results for n-Type Si Inversion Layers*, Phys. Rev. B, Vol. 5, Iss. 12, pp. 4891-4899, 1972.

[3] R.B. Fair, *Rapid Thermal Processing Science and Technology*, Academic Press, New-York, 1993.

[4] F. Roozeboom, *Advances in Rapid Thermal and Integrated Processing*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

[5] M.A. Foad, and D. Jennings, *Formation of ultra-shallow junctions by ion implantation and RTA*, Solid State Technology, pp. 43-47, 1998.

[6] Medici, User GuideVersion 2003.06, June 2003, Synopsys Inc.

[7] J. Schmitz, H.P. Tuinhout, A.H. Montree, Y.V. Ponomarev, P.A. Stolk and P.H. Woerlee, *Gate polysilicon optimization for deep-submicron MOSFETs*, ESSDERC, pp. 156-158, 1999.

[8] Y. Morimoto, Y. Jinno, K. Hirai, H. Ogata, T.Yamada and K. Yoneda, *Influence of the grain boundaries and intragrain defects on the performance of poly-Si thin film transistors*, J. Electrochem. Soc., Vol. 144(7), pp. 2495-2498, 1997.

[9] S.Saito, S. Shishigushi, K. Hamada and T. Hayashi, *Dopant profile and defect control in ion implantation by RTA with high ramp-up rate*, Mat. Chem. Phys., Vol. 54, pp. 49-53, 1998.

[10] M.I. Current, D. Lopes, M. Foad and W. Boyd, *Ultra-shallow junction technology for 100nm CMOS: xR LEAP implanter and RTP-centura rapid thermal annealer*, Mat. Chem. Phys., Vol. 54, pp.33-38, 1998.

[11] F.N. Cubaynes, P.A. Stolk, J. Verhoeven, F. Roozeboom and P.H. Woerlee, *The influence of polysilicon gate morphology on dopant activation and deactivation kinetics in deep-submicron CMOS transistors*, Mat. Science in Semi. Proc., Vol. 4, pp. 351-356, 2001.

[12] H.P. Tuinhout, A.H. Montree, J. Schmitz and P.A. Stolk, *Effects of gate depletion and boron penetration on matching of deep submicron CMOS transistors*, IEDM Techn. Digest pp. 631-33,1997.

[13] B. Yu, D.-H. Ju, W.-C. Lee, N. Kepler, T.-J. King and C. Hu, *Gate engineering for deep-submicron CMOS transistors*, IEEE Trans. Electron Dev., Vol. 45, pp. 1253-1255, 1998.

[14] S. Solmi, D. Nobili and J. Shao, *Reverse annealing, clustering, and electron mobility in arsenic doped silicon*, Journal of Applied Physics, Vol. 87, No. 2, pp. 658-662, 2000.

[15] D. Nobili, S. Solmi, M. Merli and J. Shao, *Deactivation Kinetics in Heavily Arsenic-Doped Silicon*, Journal of the Electrochemical Society, Vol. 146, No. 11, pp. 4246-4252, 1999.

[16] A. Nylandsted Larsen, B. Christensen and S. Yu. Shiryaev, *Deactivation of electricallly active arsenic in silicon during cooling-down from elevated temperatures*, Journal of Applied Physics, Vol. 71, No. 10, pp. 4854-4858, 1992.

[17] P. M. Rousseau, P. B. Griffin and J. D. Plummer, *Electrical deactivation of arsenic as a source of point defects*, Appl. Phys. Lett., Vol. 65, No. 5, pp. 578-580, 1994.

[18] C. Salm, D.T. van Veen, D.J. Gravesteijn, J. Holleman and P.H. Woerlee, *Diffusion and electrical properties of boron and arsenic doped poly-Si and poly-Ge_xSi_{1-x} ($x \approx 0.3$) as gate material for sub-0.25μm complementary metal oxide semiconductor applications*, J. Electrochem. Soc. Vol. 144(10), pp. 3665-3669, 1997.

[19] L.J. van der Pauw, *A method of measuring specific resistivity and Hall effect of discs of arbitrary shape*, Internal Philips Research report, Vol. 13, No. 1, 1958.

[20] S. Nedelec and D. Mathiot, *Kinetics of arsenic segregation at grain boundaries in polycrystalline silicon*, Semicond. Sci. Technol., Vol 12, pp. 1438-1442, 1997.

[21] TSUPREM_4, version 1998.4. Avant! Corporation, Fremont, CA, 1998.

[22] International Technology Roadmap for Semiconductors, 1999 Edition, Semiconductor Industry Association, San Jose, CA 95129.

[23] M. Rodder, Q.Z. Hong, M. Nandakumar, S. Aur, J.C. Hu and I.-C. Chen, *A sub-0.18 μm gate length CMOS technology for high performance (1.5 V) and low power (1.0 V)*, IEDM Tech. Dig., pp. 563-566, 1996.

[24] H. Park, D. Schepis, A.C. Mocuta, M. Khare, Y. Li, B. Doris, S.Shukla, T. Hugues, O. Dokumaci, S. Narasimha, S. Fung, J. Snare, B.H. Lee, J. Li, P. Ronsheim, A. Domenicucci, P. Varekamp, A.Ajmera, J. Sleight, P. O'neil, E. Maciejewski and C. Lavoie, *Gate postdoping to decouple implasnt/anneal for gate, source/drain, and extension: maximizing polysilicon gate activation for 0.1 mm CMOS technologies*, VLSI Symp. Tech. Dig., pp.134-135, 2002.

[25] J.R. Pfiester, F.K. Baker, T.C. Mele, H.-H. Tseng, P.J. Tobin, J.D. Hayden, J.W. Miller, C.D. Gunderson and L.C. Parrillo, *The effects of boron penetration on p+ polysilicon gated PMOS devices*, IEEE Trans. Electron Dev., Vol. 37(8), pp.1842-1845, 1990.

[26] J. Y. -C. Sun, C. Wong, Y. Taur and C.-H. Hsu, *Study of boron penetration through thin oxide with p+ polysilicon gate*, Dig. Int. Symp. VLSI Technology, pp. 17-19, 1989.

[27] C.G. Sodini and K.S. Krisch, …..?, IEDM Tech. Dig., pp. 617-??, 1992.

[28] G.Q. lo, W. Ting, J. Ahn and D.L. Kwong, ???, Tech. Dig. Symp. VLSI, pp. 43-??, 1991.

[29] J. Ahn, W. Ting and D.L. kwong, *Furnace nitridation of thermal SiO2 in pure N2O ambient for ULSI MOS applications*, IEEE Trans. Dev. Lett., Vol. 13, pp. 117-119, 1992.

[30] J.P. Gambino and E.G. Colgan, Materials Chemistry and Physics 52 (1998) 99.

[31] V. Misra, G. Heuss and H. Zhong, *Advanced metal electrodes for high-K dielectrics*, Proceedings of the Mat. Res. Soc. Workshop, New-Orleans, p. 5, 2000.

[32] Q. Lu, Y.C Yeo, P. Ranade, H. Takeuchi, T-J King, C. Hu, S.C. Song, H.F. Luan and D-L Kwong, *Dual-metal gate technology for deep-submicron CMOS transistors*, Dig. Int. Symp. VLSI Technology, pp. 72-73, 2000.

[33] I. De, D. Johri, A. Srivastava and C.M. Osburn, *Impact of gate workfumction on device performance at the 50 nm technology node*, Solid State Electron., Vol. 44, pp. 1077-1081, 2000.

[34] S.B. Samavedam, L.B. La, P.J. Tobin, B. White, C. Hobbs, L.R.C. Fonseca, A.A. Demkov, J. Schaeffer, E. Luckowski, A. Martinez, M. Raymond, D. Triyoso, D. Roan, V. Dhandapani, R. Garcia, S.G.H. Anderson, K. Moore, H.H. Tseng, C. Capasso, O. Adetutu, D.C. Gilmer, W.J. Taylor, R. Hegde and J. Grant, *Fermi level pinning with sub-monolayer MeOx and metal gates*, IEDM Tech. Dig., pp. 307-310, 2003.

[35] A. Agarwal, *Ultra-shallow junction formation using conventional ion implantation and rapid thermal annealing*, Axceles Technologies, Peabody, 2000.

[36] S. Saito, S. Shishiguchi, A. Mineji and T. Matsuda, *Ultra-shallow junction formation by RTA at high temperature for short heating cycle time*, Mat. Res. Soc. Symp. Proc., Vol. 532, p. 3, 1998.

[37] Y. Taur, C.H. Warm and D.J. Frank, *25 nm CMOS Design Considerations*, IEDM Tech. Dig., pp.789-792, 1998.

[38] K.K. Ng and W.Y. Lynch, *The impact of intrinsic series resistance on MOSFET scaling*, IEEE. Trans. Electron Devices, Vol. 34, pp. 503-511, 1987.

[39] K.K. Ng and W.Y. Lynch, *Analysis of gate-voltage dependent series resistance of MOSFETs, IEEE Trans. Electron devices*, Vol. 33, pp. 965-972, 1986.

[40] Y.V. Ponomarev, J.J.G.P. Loo, C.J.J. Dachs, F.N. Cubaynes, M.A. Verheijen, M. Kaiser, J.G.M. van Berkum, S. Kubicek, J. Bolk and M. Rovers, *A manufacturable 25 nm planar MOSFET technology*, Dig. Int. Symp. VLSI Technology, pp. 33-34, 2001.

[41] V. Meyssen, P.A. Stolk, J. Van Zijl, J. Van Berkum, W. Van de Wijgert, R. Lindsay, C.J.J. Dachs, G. Mannino, and N. Cowern, *Mat. Res. Soc. Symp. Proc.* 669, 2001

[42] B.J. Pawlak, R. Lindsay, R. Surdeanu, X. Pages, W. Vandervorst and K. v.d. Jeugd, *Proceedings ECS* 2003

[43] R. Lindsay, B. Pawlak, J. Kittl, K. Henson, C. Torregiani, S. Giangrandi, R. Surdeanu, W. Vandervorst, A. Mayur, J. Ross, S. McCoy, J. Gelpey, K. Elliott, X. Pages, A. Satta, A. Lauwers, P.A. Stolk, and K. Maex, *Mat. Res. Soc. Symp. Proc.* 2003

[44] C.J.J. Dachs, B.J. Pawlak, R. Surdeanu, M. van Dal, V. Venezia, G. Doornbos, R. Duffy, F. Cubaynes, C. Ravit, P.A. Stolk, R. Lindsay K. Henson, B. Dieu, L. Geenen, I. Hoflijk, O. Richard, T. Clarysse, B. Brijs, W. Vandervorst and X. Pages, *CMOS scaling beyond the 90 nm CMOS technology node shallow junction and integration challenges, USJ Workshop, proceedings?* 2003

[45] N. Cowern, B. Colombeau, R. Duffy, V.C. Venezia, C.J.J. Dachs, R. Lindsay, F. Christiano, A. Claveri, *Proceedings Mat. Res. Soc. Symp. Proc.* 2003

[46] R. Duffy, V.C. Venezia, A. Heringa, T.W.T. Husken, M.J.P. Hopstaken, N.E.B. Cowern, P.B. Griffin, C.C. Wang, *Appl. Phys. Lett.* 82 (21) 2003, p. 3647

[47] A. Mokhberi, L. Pelaz, M. Aboy, L. Marques, J. Barbolla, E. Paton, S. McCoy, J. Ross, K. Elliott, J. Gelpey, P.B. Griffin, J.D. Plummer, *IEDM Techn.Digest* 2002, p. 879

[48] H. W. Kennel, S. M. Cea, A. D. Lilak, P. H. Keys, M. D. Giles, J. Hwang, J. S. Sandford, S. Corcoran, *IEDM Techn.Digest* 2002, p. 875

[49] Bin Yu, Haihong Wang, Amol Joshi, Qi Xiang, Effiong Ibok, Ming-Ren Lin, *IEDM Techn.Digest* 2001

[50] Toshifumi Shano, Ryangsu Kim, Tetsuya Hirose, Yoshikazu Furuta, Hiroshi Tsuji, Masayuki Furuhashi, Kenji Taniguchi, *IEDM Techn.Digest* 2001

[51] R. Surdeanu, B.J. Pawlak, R. Lindsay, M. van Dal, G. Doornbos, C.J.J. Dachs, Y.V. Ponomarev, J.J.P. Loo, F.N. Cubaynes, K. Henson, M.A. Verheijen, M. Kaiser, X. Pages, P.A. Stolk, B. Taylor and M. Jurczak, *Advanced PMOS Device Architecture for Highly-Doped Ultra-Shallow Junctions*, Jap. Journal of Appl. Phys., to be published in Vol. 43, 2004.

[52] R. Lindsay, K. Henson, W. Vandervorst, K. Maex, B. J. Pawlak, R. Duffy, R. Surdeanu, P. Stolk, J. A. Kittl, S. Giangrandi, X. Pages and K. van der Jeugd, *Leakage optimization of ultra-shallow junctions formed by solid phase epitaxial regrowth*, Journal of Vacuum Science & Technology B, Vol. 22, pp. 306-311, 2004.

[53] B. J. Pawlak, R. Lindsay, R. Surdeanu, B. Dieu, L. Geenen, I. Hoflijk, O. Richard, R. Duffy, T. Clarysse, B. Brijs, W. Vandervorst and C. J. J. Dachs, *Chemical and electrical dopants profile evolution during solid phase epitaxial regrowth*, Journal of Vacuum Science & Technology B, Vol. 22, pp. 297-301, 2004.

[54] R. Surdeanu, Y.V. Ponomarev, R. Cerutti, B.J. Pawlak, L.K. Nanver, I. Hoflijk, P.A. Stolk, C.J.J. Dachs, M.A. Verheijen, M. Kaiser, M.J.P. Hopstaken, J.G.M. van Berkum, F. Roozeboom, R. Lindsay, *Laser Annealing for Ultra-Shallow Junction Formation in Advanced CMOS*, Electrochem. Soc. Symp. Proc., pp.413-418, 2002.

Chapter 7
Conclusions and Outlook

7.1 Conclusions

The aim of this work was to manufacture ultra-thin plasma nitrided oxides and integrate them in advanced CMOS transistors. The impetus for improvement has been to achieve low EOT with low gate leakage current density while maintaining high effective carrier mobility.

As presented in chapter 2, the power consumption for future integrated circuits could dramatically increase if the gate leakage current is not reduced. Consequently, the gain in performance from scaling the gate oxide might not be worth the problems associated with increased power consumption. This is especially true for portable applications that drastically limit the maximum allowed gate leakage current in order to extend the lifetime of the battery. To reduce the gate leakage current and continue the scaling of the thickness, pure oxide gate dielectrics have been replaced by higher permittivity dielectrics.

In this work, plasma nitrided oxides have been investigated to replace pure oxide and lightly nitrided oxide gate dielectrics. A significant decrease of a factor 10 in gate leakage current is observed for MOS devices with a plasma nitrided oxide as compared to gate oxide. Yet, ultra-thin plasma nitrided oxide films are required in advanced CMOS technologies and the goal of this thesis work was to investigate the scalability of such dielectrics and their integration into advanced CMOS process flows.
The first important question addressed in this work was how to characterize ultra-thin gate dielectric films. In chapter 3, the capacitance-voltage (C-V) measurement methodology has been reviewed for devices with a high gate leakage current (i.e. having an ultra-thin plasma nitrided gate oxide). It was shown that high frequency C-V measurements (up to 1 MHz) under high gate leakage current yield large measurement errors and cannot be used for the extraction of important parameters such as the EOT, the electrical thickness, the polysilicon and substrate doping

concentrations. These errors can be minimized by optimizing the test structure until a certain gate leakage current density that was estimated to about 100 A/cm^2 for an ideal like test structure. An RF (400 MHz – 1 GHz) C-V measurement methodology has been proposed as an alternative to the HF measurements. Accurate C-V characteristics of devices with very high gate leakage current (the method has been proven up to a gate leakage current density of 1000 A/cm^2) have been successfully obtained.

In chapter 4, the physical characterization of ultra-thin plasma nitrided oxide has been studied. The very small thicknesses of the plasma nitrided oxides as well as the hybrid nature of these films represent an incredible challenge for the material analysis. Techniques enabling the measurement of the thickness, N distribution and N content have been optimized. A benchmark of these various techniques have been also made. Finally, the N incorporation mechanism in ultra-thin plasma nitrided films has been investigated. It was found that N atoms seem to exchange exclusively with O atoms in the SiO_2 matrix and form strong Si≡N bonds.

The characterization techniques, selected in chapters 3 and 4, have been used to optimize ultra-thin plasma nitrided oxides. The base oxide, plasma nitridation and post nitridation anneal processing steps have been optimized. It was found that scaling aggressively the base oxide thickness does not yield necessarily to the best leakage current and performance trade-off. However, decreasing the energy of the plasma while increasing the density of N ions and neutrals in the plasma has resulted in higher channel mobility and possibly better scalability of the gate dielectric. This "soft" plasma has been achieved by pulsing the RF source power and maximizing the off-time per cycle. The role of the PNA has been identified as a stabilization step of the SiON film. Because it is processed at high temperature, it also cures defects that could have been created during the plasma step. Finally, the benefit of performing the PNA in an oxidizing ambient for improved mobility has to be traded-off with the scalability of the layer.

The intrinsic reliability of the optimized ultra-thin plasma nitrided oxides has been measured. While a lower acceleration factor was obtained for plasma nitrided oxides relative to pure oxide, it is believed that the intrinsic reliability of such films will not be a showstopper for their integration into circuits. It is believed that the limited factor will be the gate leakage current.

This thesis work ends with the integration of the optimized ultra-thin plasma nitrided oxide in advanced CMOS transistors. The gate stack has been optimized with a view to maximize the gate activation and reduced depletion in the polysilicon electrode. It was shown that the polysilicon gate activation can be enhanced by adding a pre-doping implantation and optimizing the thermal budget of the CMOS process flow. The compatibility of such gate stack with advanced ultra-shallow junctions formed by pre-amorphization and annealed with a conventional RTA or using a SPER process, has been addressed. There is a trade-off between getting a fully activated polysilicon gate that requires high temperature anneal and the formation of abrupt shallow junctions. It was found that gate leakage current and polysilicon depletion can be increased as a result of the formation of abrupt shallow junctions. The introduction of metallic gates would eliminate polysilicon depletion effects and will be compatible with low thermal budget junctions formation processes. However,

many issues need to be solved before their integration in a CMOS process flow. For bulk Si technology, the use of two metal electrodes with the required band edge work functions to obtain low symmetrical threshold voltages in N- and PMOS devices remains an incredible challenge. At the time of this thesis is written, no metals have been identified as N- or PMOS solution.

7.2 Outlook

There are two main questions that still need to be answered:

1. Can plasma nitrided oxide still be used for future CMOS technology nodes?
2. Will CMOS downscaling continue for many technology nodes as presented in the ITRS (so far till 16 nm L_G)?

Using conventional planar bulk Si technology, ultra-thin plasma nitrided oxide could be still used for high-performance applications (such as microprocessors) where extremely thin films are required and where the gate leakage specification is not aggressive.
However, for portable applications that required very low power consumption, plasma nitrided oxide will reach very quickly its scalability limit. Some technological changes such as the use of strained-Si or the implementation of metal gates might help in relaxing the thickness of the gate dielectric and therefore meet the aggressive gate leakage current requirement. However, such technological changes will only be the solution for one CMOS generation node. High-*K* dielectrics will be required to decrease the leakage and therefore power consumption of such a chip. However, at the time this thesis is written, significant issues remain such as the "pinning" of the Fermi-level when used in combination with polysilicon electrodes, high fixed (bulk) charge, high interface state density and V_T instability due to transient charge trapping and low effective mobility.

Using multigate architectures such as Finfet could relax the thickness requirement of the gate dielectric. Indeed, the very large performance improvement obtained with these transistors could allow thicker gate dielectric yielding a significant reduction of the gate leakage current. Also, a decrease of V_{DD} could yield a small decrease of the gate leakage current but will have a larger impact on the reduction of the active power.

If no breakthrough is found in the high-K gate dielectric world to achieve thicknesses in the atomic scale, then the CMOS downscaling will stop. At this moment, high-*K* films require the use of an interfacial oxide or oxynitride layer that limit the scalability of the dielectric. Alternatives are currently under investigation such as molecular transistors. The main challenge for these new concepts is to be ready in time to continue the race towards smaller transistors.

Appendix A

A.1 Equivalent Oxide Thickness (EOT)

The EOT is the thickness of a SiO_2 film having the same specific capacitance as the dielectric film in question (A.1).

$$EOT = \frac{\varepsilon_{SiO_2} T_{physical}}{\varepsilon_{dielectric}} \tag{A.1}$$

where EOT and $T_{physical}$ are the equivalent oxide and physical thickness, respectively, and ε_{SiO2} and $\varepsilon_{dielectric}$ are respectively the dielectric constant of silicon dioxide and that of the dielectric studied.

Because of its independence of device structural effects, the EOT is basically a materials parameter. It is normally evaluated with the device biased in weak accumulation, where errors involved in its calculation are expected to be minimized. Care should be taken when comparing the EOT calculated from different simulators/models as significant differences have been noticed. It is wise to choose one of the models, and always specify it when quoting values (see section 5 of chapter 4)

A.2 Capacitance Equivalent Thickness (CET)

This thickness is simply measured on the C-V curve (A.2). All device-related shortcomings are included in the CET parameter. It does, however, govern device operating parameters such as drive current. Thus it is usually evaluated in inversion, the operating mode of the device, yielding the parameter CET_{inv}.

$$CET = \frac{\varepsilon_{SiO_2} \varepsilon_0 A}{C_{meas}} \tag{A.2}$$

where ε_{SiO2} is the oxide permittivity, ε_0 is the permittivity in vacuum, A is the area of the test structure and C_{meas} is the measured capacitance.