Integration of thulium silicate for enhanced scalability of high-k/metal gate CMOS technology

EUGENIO DENTONI LITTA

Doctoral Thesis in Information and Communication Technology
Stockholm, Sweden 2014
A dissertation submitted to KTH Royal Institute of Technology, Stockholm, Sweden, in partial fulfillment of the requirements for the degree of Teknologe Doktor (Doctor of Philosophy). The public defense will take place on 27 May 2014 at 10.00 a.m. in Sal D, Forum, Kungliga Tekniska Högskolan, Isafjordsgatan 39, Kista.

© Eugenio Dentoni Litta, May 2014

Printed by Universitetsservice US-AB

Cover image: Illustration of the workflow of the thesis. The development of a novel ALD process (left) has enabled the fabrication of TmSiO/HfO₂ MOSFETs (center) demonstrating enhanced EOT-mobility trade-off over conventional SiOₓ/HfO₂ devices (right).
Abstract

High-k/metal gate stacks have been introduced in CMOS technology during the last decade in order to sustain continued device scaling and ever-improving circuit performance. Starting from the 45 nm technology node, the stringent requirements in terms of equivalent oxide thickness and gate current density have rendered the replacement of the conventional SiON/poly-Si stack unavoidable. Although Hf-based technology has become the de facto industry standard for high-k/metal gate MOSFETs, problematic long-term scalability has motivated the research of novel materials and solutions to fulfill the target performances expected of gate stacks in future technology nodes.

In this work, integration of a high-k interfacial layer has been identified as the most promising approach to improve gate dielectric scalability, since this technology presents the advantage of potential compatibility with both current Hf-based and plausible future higher-k materials. Thulium silicate has been selected as candidate material for integration as interfacial layer, thanks to its unique properties which enabled the development of a straightforward integration process achieving well-controlled and repeatable growth in the sub-nm thickness regime, a contribution of \((0.25 \pm 0.15)\) nm to the total EOT, and high quality of the interface with Si.

Compatibility with industry-standard CMOS integration flows has been kept as a top priority in the development of the new technology. To this aim, a novel ALD process has been developed and characterized, and a manufacturable process flow for integration of thulium silicate in a generic gate stack has been designed. The thulium silicate interfacial layer technology has been verified to be compatible with standard integration flows, and fabrication of high-k/metal gate MOSFETs with excellent electrical characteristics has been demonstrated.

The possibility to achieve high performance devices by integration of thulium silicate in current Hf-based technology has been specifically demonstrated, and the TmSiO/HfO\(_2\) dielectric stack has been shown to be compatible with the industrial requirements of operation in the sub-nm EOT range (down to 0.6 nm), reliable device operation over a 10 year expected lifetime, and compatibility with common threshold voltage control techniques. The thulium silicate interfacial layer technology has been especially demonstrated to be superior to conventional chemical oxidation in terms of channel mobility at sub-nm EOT, since the TmSiO/HfO\(_2\) dielectric stack achieved \(~20\%\) higher electron and hole mobility compared to state-of-the-art SiO\(_x\)/HfO\(_2\) devices at the same EOT. Such performance enhancement can provide a strong advantage in the EOT-mobility trade-off which is commonly observed in scaled gate stacks, and has been linked by temperature and stress analyses to the higher physical thickness of the high-k interfacial layer, which results in attenuated remote phonon scattering compared to a SiO\(_x\) interfacial layer achieving the same EOT.

**Keywords:** thulium, silicate, TmSiO, Tm2O3, interfacial layer, IL, CMOS, high-k, ALD
# Table of Contents

Acknowledgments vi  
List of appended papers vii  
Related work not included in the thesis viii  
Summary of appended papers x  
List of Acronyms and Symbols xiii  
Introduction 1  

1 State of the art in high-k/metal gate technology 5  
1.1 High-k/metal gate technology in the manufacturing environment . 5  
1.1.1 Gate-first integration scheme . . . . . . . . . . . . . . . . . . . . . . . 7  
1.1.2 Gate-last integration scheme . . . . . . . . . . . . . . . . . . . . . . . 8  
1.2 Effect of high-k dielectrics on device properties . . . . . . . . . . . . . . . 9  
1.2.1 Equivalent oxide thickness . . . . . . . . . . . . . . . . . . . . . . . . . 10  
1.2.2 Channel mobility . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11  
1.2.3 Threshold voltage . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14  
1.2.4 Reliability . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16  
1.3 Scalability issues in SiO$_x$/HfO$_2$ stacks . . . . . . . . . . . . . . . . . . . . . 16  
1.3.1 Scaling of the interfacial layer . . . . . . . . . . . . . . . . . . . . . . . . . 17  
1.4 Research directions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19  
1.4.1 Higher-k dielectrics . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20  
1.4.2 High-k interfacial layers . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 21  

2 Integration of TmSiO as high-k interfacial layer 23  
2.1 Design guidelines for a high-k interfacial layer . . . . . . . . . . . . . . . . . . . . 23  
2.1.1 Design considerations . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23  
2.1.2 Choice of material . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 25  
2.2 Integration of TmSiO in a CMOS process . . . . . . . . . . . . . . . . . . . . . . . . . 26  
2.2.1 Surface cleaning . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 27  
2.2.2 Deposition of Tm$_2$O$_3$ . . . . . . . . . . . . . . . . . . . . . . . . . . . . 28  
2.2.3 Formation of thulium silicate . . . . . . . . . . . . . . . . . . . . . . . . . 28  
2.2.4 Selective removal of Tm$_2$O$_3$ . . . . . . . . . . . . . . . . . . . . . . . . 30  
2.2.5 Removal of TmSiO . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 31  
2.3 Compatibility with standard CMOS process flows . . . . . . . . . . . . . . . . . . . . 31  
2.3.1 Gate-last integration scheme . . . . . . . . . . . . . . . . . . . . . . . . . . 32
Acknowledgments

This thesis summarizes my work for the last four years as a Ph.D. student at KTH, where I have worked with many people to whom I would like to express my gratitude.

First of all, I would like to thank my supervisor, Docent Per-Erik Hellström, who has always advised and encouraged me in my research. His broad knowledge in device physics and his strong experience in semiconductor processing have been invaluable; I have learned a lot during our meetings. Prof. Mikael Östling, who has been my co-supervisor during these years, I want to thank for always supporting my work. He has been a constant source of inspiration for me.

I am also grateful to Prof. Carl-Mikael Zetterling, Prof. Anders Hallén, Prof. Gunnar Malm, Prof. Max Lemme and Docent Henry Radamson for valuable help and useful discussions. I also want to thank Christian Ridder and Sven Valerio for their help in the cleanroom and for the many interesting conversations, both work-related and non.

I have been lucky in meeting nice office roommates in Dr. Benedetto Buono, Dr. Jinghua Xia and Shuoben Hou. I have also enjoyed the company and cleanroom help of many people, especially Dr. Yong-Bin Wang, Gabriel Roupillard, Timo Söderqvist, Dr. Christoph Henkel, Dr. Thomas Zabel, Maziar Naiini, Luigia Lanni, Ganesh Jayakumar, Konstantinos Garidis, Ali Asadollahi and Valentine Dubois.

I want to thank everyone at the department of Integrated Devices and Circuits for creating a very nice working environment: Ahmad Abedin, Jesper Berggren, Anders Eklund, Hossein Elahipanah, Gunilla Gabrielsson, Dr. Valur Gudmundsson, Dr. Oscar Gustafsson, Prof. Mattias Hammar, Raheleh Hedayati, Saleh Kargarrazi, Dr. Jiantong Li, Dr. Margareta Linnarsson, Dr. David Martin, Maryam Olyaei, Carl Reuterskiöld Hedlund, Prof. Ana Rusu, Arash Salemi, Katarina Smedfors, Anderson Smith, Sethu Saveda Suwanam, Ye Tian, Dr. Muhammad Usman and Sam Vaziri.

Finally, none of this would have been possible without continuous support and encouragement from my family and from my girlfriend, Maria Rosaria, whom I can not thank enough.

Eugenio Dentoni Litta
Stockholm, April 2014
List of appended papers

I  In situ SiO$_x$ interfacial layer formation for scaled ALD high-k/metal gate stacks
   E. Dentoni Litta, P.-E. Hellström, C. Henkel, M. Östling
   13th International Conference on Ultimate Integration on Silicon (ULIS), pp. 105-108, 2012

II Characterization of thulium silicate interfacial layer for high-k/metal gate MOSFETs
   E. Dentoni Litta, P.-E. Hellström, C. Henkel, M. Östling
   14th International Conference on Ultimate Integration on Silicon (ULIS), pp. 121-124, 2013

III High-deposition-rate atomic layer deposition of thulium oxide from TmCp$_3$ and H$_2$O
   E. Dentoni Litta, P.-E. Hellström, C. Henkel, S. Valerio, A. Hallén, M. Östling

IV Thulium silicate interfacial layer for scalable high-k/metal gate stacks
   E. Dentoni Litta, P.-E. Hellström, C. Henkel, M. Östling

V Effective workfunction control in TmSiO/HfO$_2$ high-k/metal gate stacks
   E. Dentoni Litta, P.-E. Hellström, M. Östling
   15th International Conference on Ultimate Integration on Silicon (ULIS), 2014

VI Mobility enhancement by integration of TmSiO IL in 0.65nm EOT high-k/metal gate MOSFETs
   E. Dentoni Litta, P.-E. Hellström, M. Östling
   European Solid-State Device Research Conference (ESSDERC), 2013

VII Enhanced channel mobility by integration of sub-nm-EOT TmSiO/HfO$_2$/TiN high-k/metal gate MOSFETs
   E. Dentoni Litta, P.-E. Hellström, M. Östling
   In manuscript
Related work not included in the thesis


4 E. Dentoni Litta, P.-E. Hellström, C. Henkel, and M. Östling, “Electrical characterization of thulium silicate interfacial layers for integration in high-k/metal gate CMOS technology,” *Solid-State Electronics*, accepted for publication


Summary of appended papers

Paper I. In situ SiO$_x$ interfacial layer formation for scaled ALD high-k/metal gate stacks. This paper describes the development of a method for *in situ* oxidation of the Si substrate as a way to form a chemical oxide interfacial layer before deposition of the bulk high-k layer. Integration in HfO$_2$-based MOSFETs shows that the method achieves comparable quality to conventional low-temperature wet chemical oxidation. The author has performed 90% of the experimental design, 100% of the fabrication, 100% of the characterization and data analysis and 90% of the manuscript writing.

Paper II. Characterization of thulium silicate interfacial layer for high-k/metal gate MOSFETs. This paper evaluates Tm and La silicates for integration as high-k interfacial layers in generic gate stacks. TmSiO is shown to provide superior electrical properties and to be compatible with standard gate-last and gate-first integration flows. The author has performed 100% of the experimental design, 100% of the fabrication, 100% of the characterization and data analysis and 90% of the manuscript writing.

Paper III. High-deposition-rate atomic layer deposition of thulium oxide from TmC$_3$P$_3$ and H$_2$O. This paper documents the development of a novel ALD process for deposition of Tm$_2$O$_3$ and reports a thorough characterization of the properties of the deposited films. The process is shown to be advantageous over the pre-existing deposition process in terms of a higher deposition rate and of the capability to deposit Tm$_2$O$_3$ in direct contact with the underlying Si. The author has performed 100% of the process design and development, 80% of the characterization and data analysis and 80% of the manuscript writing.

Paper IV. Thulium silicate interfacial layer for scalable high-k/metal gate stacks. This paper reports on the integration of thulium silicate as interfacial layer in a high-k/metal gate CMOS process. Evaluation of the resulting device properties evidences that TmSiO is competitive toward conventional chemical oxide, especially in terms of IL EOT and channel mobility. The author has performed 100% of the experimental design, 100% of the
fabrication, 100% of the characterization and data analysis and 90% of the manuscript writing.

**Paper V. Effective workfunction control in TmSiO/HfO₂ high-k/metal gate stacks.** This paper evaluates the effect of the integration of thulium silicate interfacial layers on the capability to properly set the threshold voltages for nFET and pFET operation. Compatibility with commonly employed effective workfunction tuning techniques, which rely on the implementation of a dual metal or a dual capping layer process, is demonstrated for the TmSiO/HfO₂ dielectric stack. The author has performed 100% of the experimental design, 100% of the fabrication, 100% of the characterization and data analysis and 90% of the manuscript writing.

**Paper VI. Mobility enhancement by integration of TmSiO IL in 0.65nm EOT high-k/metal gate MOSFETs.** This paper reports on the channel mobility enhancement which can be achieved by integration of thulium silicate as a replacement of conventional chemical oxide in sub-nm-EOT high-k/metal gate MOSFETs. The TmSiO/HfO₂ dielectric stack is shown to outperform state-of-the-art SiOₓ/HfO₂ devices at EOT as low as 0.65 nm in terms of 15-20% higher electron and hole mobility. The author has performed 100% of the experimental design, 100% of the fabrication, 100% of the characterization and data analysis and 90% of the manuscript writing.

**Paper VII. Enhanced channel mobility by integration of sub-nm-EOT TmSiO/HfO₂/TiN high-k/metal gate MOSFETs.** This paper discusses optimized annealing conditions for the TmSiO/HfO₂ stack, which lead to competitive gate leakage current density and reliable 10 year device operation in terms of oxide integrity and threshold voltage instability, at sub-nm EOT (as low as 0.6 nm) and 15-20% higher electron and hole mobility than state-of-the-art SiOₓ/HfO₂ devices. The author has performed 100% of the experimental design, 100% of the fabrication, 100% of the characterization and data analysis and 90% of the manuscript writing.
List of Acronyms and Symbols

AC ....... alternating current
ALCVD ...... atomic layer chemical vapor deposition
ALD ...... atomic layer deposition
ALE ...... atomic layer epitaxy
BEOL ...... back end of line
BTI ...... bias-temperature instability
CET ...... capacitance equivalent thickness
$C_{GB}$ ...... gate-to-bulk capacitance-voltage characteristic
$C_{GC}$ ...... gate-to-channel capacitance-voltage characteristic
$C_{\text{inv}}$ ...... gate capacitance density in inversion
CMOS ...... complementary metal-oxide-semiconductor
CMP ...... chemical-mechanical polishing
$C_{\text{ox}}$ ...... areal capacitance density of the oxide
$C_p$ ...... cyclopentadienyl
CV ...... capacitance-voltage
CVD ...... chemical vapor deposition
CVS ...... constant voltage stress
DC ...... direct current
DIBL ...... drain-induced barrier lowering
$D_{it}$ ...... interface state density
DRAM . . . . dynamic random access memory

$E_{\text{eff}}$ . . . . effective field

EOT . . . . . equivalent oxide thickness

EWF . . . . . effective work function

FEOL . . . . front end of line

FGA . . . . . forming gas anneal

GPC . . . . . growth per cycle

IC . . . . . . integrated circuit

$I_D-V_G$ . . . . drain current versus gate voltage MOSFET IV characteristics

IL . . . . . . interfacial layer

$I_{\text{OFF}}$ . . . . off-state current

$I_{\text{ON}}$ . . . . drive current

ITRS . . . . . International Technology Roadmap for Semiconductors

IV . . . . . . current-voltage

$L_G$ . . . . . gate length

LOCOS . . . localized oxidation of silicon

$\mu$ . . . . . . channel mobility

MIPS . . . . . metal-inserted poly-silicon

MOL . . . . . middle of line

MOSFET . . . metal-oxide-semiconductor field-effect transistor

MSM . . . . . measure-stress-measure

MuGFET . . multiple gate field-effect transistor

NBTI . . . . . negative bias-temperature instability

$N_{\text{inv}}$ . . . . inversion charge density

PBTI . . . . . positive bias-temperature instability

PDA . . . . . post deposition anneal

PEALD . . . plasma-enhanced atomic layer deposition
PECVD . . . plasma-enhanced chemical vapor deposition
PMA . . . . . . post metallization anneal
PVD . . . . . physical vapor deposition
RBS . . . . . Rutherford backscattering spectroscopy
RCS . . . . . remote Coulomb scattering
RGA . . . . . residual gas analysis
RMG . . . . . replacement metal gate
RPS . . . . . remote phonon scattering
RSRS . . . . remote surface roughness scattering
RTA . . . . . rapid thermal anneal
SC-1 . . . . . standard clean 1
SC-2 . . . . . standard clean 2
SCE . . . . . short-channel effect
SOI . . . . . silicon on insulator
SQL . . . . . structured query language
STI . . . . . shallow trench isolation
TDDB . . . . time-dependent dielectric breakdown
TEM . . . . . transmission electron microscopy
thd . . . . . 2,2,6,6-tetramethyl-3,5-heptanedione
Tof-ERDA . . . time of flight elastic recoil detection analysis
V_{DD} . . . . . supply voltage
VLSI . . . . . very-large-scale integration
VRS . . . . . voltage ramp stress
V_{T} . . . . . threshold voltage
W_{G} . . . . . gate width
Introduction

Complementary metal-oxide-semiconductor (CMOS) technology has underpinned the great advances achieved by the microelectronics industry in the last decades, which have led to today’s wide availability of personal computing devices and formed the basis of the digital revolution. The capability to continuously provide increased functionality per area, power and cost is central to the evolution of digital electronics and to the development of new technological realities such as the Internet of Things.

Geometrical scaling of the metal-oxide-semiconductor field-effect transistor (MOSFET) is the central concept which allowed the exponential growth in devices integrated on a single chip known as Moore’s Law. The number of transistors on a chip has increased from thousands in the first microprocessors to a few billions in state-of-the-art very-large-scale integration (VLSI) integrated circuits (ICs), at constantly reduced cost and power dissipation per device and constantly increased performance. Simple geometrical scaling has allowed this outstanding evolution for decades without substantial change to the device structure and to the materials involved. Since the 90 nm node, however, the introduction of novel materials and structures has been needed in order to achieve the target performance improvements or in order to overcome physical limits of the conventional materials.

One crucial and disruptive innovation has been the introduction of high-k/metal gates in place of the conventional SiON/poly-Si gate stack, rendered necessary by the conflicting requirements in terms of gate capacitance density and gate current density. The extensive research on high-k dielectrics carried out during the late 1990’s and early 2000’s has led to the introduction of high-k/metal gate technology in manufacturing, first announced at the 45 nm node by Intel and IBM [1, 2]. Hf-based high-k/metal gate technology is integrated in virtually all of today’s 22–20 nm technologies [3–5] and will most likely be at the core of the forthcoming 16 and 14 nm technologies. However, the 2013 Edition of the International Technology Roadmap for Semiconductors (ITRS) has identified the integration of novel high-k/metal gate technologies as needed in the short term, since current technological solutions are predicted to face strong challenges in fulfilling the requirements set for High Performance and Low Standby Power processes as early as 2016 [6].

Research on future high-k dielectrics has focused in two directions: on one hand, replacement of Hf-based dielectrics with novel materials possessing higher dielectric constant; on the other hand, integration of a high-k interfacial layer in place of the
Introduction

commonly employed chemical oxide or oxynitride. The second approach presents the advantage of preserving compatibility with existing Hf-based technology while at the same time being potentially adaptable to integration of future higher-k dielectrics.

In this experimental work, thulium silicate has been identified as a promising high-k interfacial layer, thanks to its favorable material properties which enabled the development of a straightforward and manufacturable integration process with controllable interfacial layer thickness in the sub-nm range. Compatibility with standard CMOS processes and requirements is a crucial prerequisite for the introduction of new materials, therefore a strong emphasis has been put on evaluating thulium silicate from the point of view of integration in a manufacturable CMOS process. This viewpoint has driven the preliminary phase of literature study to indentify suitable candidates, as well as the development phase for the integration process flow. The same perspective has also motivated the development and characterization of a novel atomic layer deposition (ALD) process, whose design principles and implementation would ensure a well-controllable interfacial layer process and its future applicability to three-dimensional device structures, such as FinFETs and nanowires.

Fabrication and characterization of high-k/metal gate MOSFETs has been at the core of this work, since integration in an actual CMOS process is crucial in assessing the compatibility of a new gate stack with critical device properties. Besides gate capacitance scaling and gate current density, which constitute the primary motivation for integration of high-k dielectrics, qualification of the novel interfacial layer has been performed through a quantitative analysis of its effects on the main MOSFET device properties, especially threshold voltage, channel mobility, interface quality and reliability. Even though the process has been developed for straightforward integration in a generic gate stack, compatibility with industrially relevant HfO$_2$-based gate stacks has been specifically addressed, leading to the conclusion that the TmSiO interfacial layer technology is indeed a viable solution for future CMOS technology nodes.

This thesis is organized as follow.

Chapter 1 reviews the state of the art in high-k/metal gate CMOS technology and details the effects of the integration of high-k dielectrics on MOSFET device properties. The scalability issues that have been identified in the current Hf-based technology are then described, followed by a presentation of possible solutions and research directions.

Chapter 2 motivates the selection of thulium silicate as a candidate high-k interfacial layer and describes the development of a CMOS-compatible integration process flow. The compatibility of the material with common gate-first and gate-last integration schemes is also addressed.

Chapter 3 describes the principles of atomic layer deposition of high-k gate dielectrics and reports the development of a novel ALD process for Tm$_2$O$_3$. The resulting film properties are analyzed and their relevance to the integration of thulium silicate as interfacial layer is discussed.

Chapter 4 describes the process flow employed in this work to fabricate high-
Introduction

k/metal gate MOSFETs and performs a comparative analysis of MOSFETs using thulium silicate and conventional chemical oxide as interfacial layers. Compatibility of the TmSiO/HfO$_2$ gate dielectric stack with commonly employed threshold voltage control techniques is then evaluated, followed by the description of a preliminary experiment aimed at co-integrating the high-k interfacial layer with a higher-k bulk oxide.

Chapter 5 describes the integration of thulium silicate in scaled Hf-based gate stacks, achieving improved channel mobility compared to state-of-the-art devices using chemical oxide as interfacial layer. The development of optimized annealing conditions for the TmSiO/HfO$_2$ dielectric stack is also documented, together with their effect on obtaining competitive gate current density and reliable 10-year device operation. The observed enhancement in channel mobility is then analyzed in detail, via measurements at varying temperature and under voltage stress.

Finally, the main conclusions from this work are presented and an outlook on possible future developments is given.
CMOS scaling has entered a non-classical era, where the standard geometrical scaling has to be complemented by the introduction of novel materials and structures in order to achieve the target performance goals. A crucial step in this path has been the introduction of high-k/metal gate technology, a disruptive but necessary step in order to continue with gate capacitance scaling.

This chapter briefly reviews the state of the art in high-k technology for CMOS logic. Given that high-k/metal gate is a mature manufacturing technology, its current industrial implementations are described first, in a way that will also serve as reference for the rest of the discussion, since compatibility to standard CMOS processes was considered as a priority throughout the work. The multiple effects of the introduction of a novel gate stack on device performance are then discussed, followed by a description of the scalability issues of the current SiO$_x$/HfO$_2$ technology. Finally, a short review of the main research directions addressed in literature is presented. The chapter also serves to introduce the terminology and concepts relevant to the discussion presented in this thesis.

1.1 High-k/metal gate technology in the manufacturing environment

CMOS scaling has followed for decades the geometrical rules first stated by Dennard [7] and later generalized in different forms of scaling [8]. In general, device scaling requires systematic variations in geometry, voltage and doping, with the aim of achieving an electrical behavior similar to long-channel devices even at short gate lengths. Scaling rules are specifically designed to suppress short-channel effects (SCEs), which, in the common modern use of the term, represent the overall effect of the lateral (i.e. parallel to the current flow in the channel) electric field distribution on the electrical characteristics of scaled devices. SCEs cause an unwanted depen-
1. State of the art in high-k/metal gate technology

dence of threshold voltage on gate length and on drain bias, due to depletion of part of the channel by the PN junctions at the source/drain and due to the influence of drain bias on the energy barrier to the injection of carriers from the source into the channel, known as drain-induced barrier lowering (DIBL). Stringent control of the threshold voltage is required in any CMOS process, since it determines the trade-off between high $I_{ON}$ (maximum drive current, measured when the gate and drain terminals are connected to the supply voltage $V_{DD}$) and low $I_{OFF}$ (leakage current when the device is off, i.e. at 0 V gate bias). Additionally, symmetric nFET and pFET threshold voltages are desirable in digital electronics in order to achieve symmetric noise margins. SCEs, in the form of DIBL and punchthrough (i.e. the merging of the source and drain depletion regions), also result in increased $I_{OFF}$, requiring the implementation of additional process steps and complex bi-dimensional doping profiles to achieve acceptable off-state leakage current. Punchthrough is especially undesired, since it results in a strong loss of control of the gate over the subthreshold current, visible in the MOSFET $I_D$-$V_G$ characteristics as an increased subthreshold slope.

Despite the increasing difficulties in suppressing SCEs and achieving low off-state power dissipation, MOSFET scaling has continued for decades because of the benefits of higher performance, lower cost per device and higher packing density. Gate oxide scaling is a crucial part of the generalized device scaling rules, since it results in increased control of the gate over the channel. It also has a benefic effect on drive current, although this is partly counteracted in deeply scaled devices by mobility degradation (see Section 1.2.2). For decades, thermally grown SiO$_2$ has sustained continuous gate oxide scaling in MOSFETs, until increasing gate leakage current and reliability concerns in films thinner than $\sim$3 nm led to the development of nitridation techniques and to the resulting integration of SiON as gate dielectric. SiON is advantageous over SiO$_2$ for very thin dielectric films in terms of improved reliability, reduced impurity diffusion and slightly higher dielectric constant, thereby achieving the same capacitance density ($C_{ox}$) at higher physical thickness. However, a barrier to SiON thickness scaling was hit at $\sim$1.2 nm due to the exponentially increasing direct tunneling current, at which point high-k dielectrics had to be integrated. Such materials, having higher dielectric constant than SiON, have allowed continued scaling of $C_{ox}$ to the present day.

The introduction of high-k dielectrics in manufacturing has required the concurrent integration of metal gate technology, because of the observed incompatibility of high-k dielectrics with conventional poly-Si gate electrodes, resulting in thermodynamic instability, difficulties in setting the threshold voltage and mobility degradation [9]. Therefore, the two technologies are usually considered together and are referred to as high-k/metal gate technology. The reduced thermodynamic stability of the high-k/metal system, compared to the SiON/poly-Si system, has led to the development of two different integration schemes for high-k/metal gate technology in CMOS processing, i.e. the gate-first and gate-last integration schemes (Fig. 1.1).
1.1. High-k/metal gate technology in the manufacturing environment

Figure 1.1. (a) Schematic view of the high-k/metal gate stack in a gate-first integration scheme and (b) cross-section of a gate-first MOSFET manufactured by IBM (reproduced from [10]). (c) Schematic view of the high-k/metal gate stack in a gate-last integration scheme and (d) cross-section of a gate-last MOSFET manufactured by Intel (reproduced from [11]).

1.1.1 Gate-first integration scheme

The gate-first integration scheme is the simplest from the processing point of view, in that it preserves maximum compatibility with the standard CMOS process but succeeds in replacing the conventional SiON/poly-Si gate stack with a high-k/metal gate stack providing acceptable device properties [12].

The process resembles a standard SiON/poly-Si process, except for the gate formation module. The SiON dielectric is replaced by a high-k dielectric, usually deposited on top of a thin (~1 nm or less) interfacial layer, which is formed by chemical oxidation at low temperature (leading to a non-stoichiometric SiO$_x$), possibly followed by a nitridation step. Regarding the high-k dielectric itself, Hf-based dielectrics are the de facto standard in production. In the case of gate-first processing,
1. State of the art in high-k/metal gate technology

common examples are HfO$_2$ and HfSiON, with the former having the advantage of a higher dielectric constant while the latter provides higher thermodynamic stability. After the high-k dielectric deposition, which is typically performed by ALD (see Chapter 3), a dielectric capping layer can be integrated to facilitate threshold voltage control. A dual cap process (e.g. La$_2$O$_3$ for nFETs and Al$_2$O$_3$ for pFETs), together with a single mid-gap metal, has been shown to provide the desired symmetric device operation [13]. The gate electrode in gate-first processing consists in a metal-inserted poly-silicon (MIPS) stack, which uses a single metal gate (e.g. TiN) as workfunction-setting metal and a conventional poly-Si layer to maintain compatibility with self-aligned source/drain implantation and silicidation. A scavenging element can be integrated in the MIPS stack with the purpose of $C_{ox}$ scaling (see Section 1.3), while the poly-Si layer also serves as an oxygen diffusion barrier to minimize oxidation of the substrate and the gate metal during dopant activation annealing.

The advantage of the gate-first integration scheme lies in its strong similarity to a conventional SiON/poly-Si CMOS process, thereby requiring a contained process development effort and a limited number of additional process steps (e.g. only one extra lithography step is required to implement a dual cap process). However, the high-k/metal gate stack is exposed to the full thermal budget for dopant activation, which has serious consequences on the resulting device properties and requires careful development and optimization of the gate stack composition and of the annealing conditions. The main issue is a substantial shift of the effective work function (EWF) of the stack after annealing. While this phenomenon can be somewhat balanced by the integration of capping layers, it renders low pFET threshold voltage extremely difficult to achieve at low EOT (see Section 1.2.3).

1.1.2 Gate-last integration scheme

The gate-last integration scheme, also known as replacement metal gate (RMG), has been developed with the purpose of minimizing degradation of the high-k/metal gate stack by exposing it to a limited thermal budget. To this aim, a dummy SiO$_2$/poly-Si gate stack is used during the front end of line (FEOL) and part of the middle of line (MOL), i.e. for the entire fabrication of MOSFETs up to dopant activation, silicidation and deposition of the first interlayer dielectric. The dielectric is then planarized via chemical-mechanical polishing (CMP), the exposed dummy gate is etched and the high-k/metal gate stack is deposited in the gate trench, followed by a low-resistivity gate-fill metal (e.g. Al or W). After planarization of the gate-fill metal by CMP, the process continues with contact vias and back end of line (BEOL) metal interconnections as in a conventional process flow [8, 12].

The most commonly used gate dielectric in gate-last process flows is HfO$_2$, typically deposited on a sub-nm SiO(N) layer, obtained by chemical oxidation of the Si substrate and optional nitridation. The main advantage of the gate-last scheme is that the EWF of the stack is effectively set by the vacuum workfunction of the metal, thanks to the reduced thermal budget. A dual-metal process without dielectric
capping layers is typically employed to obtain symmetric threshold voltages, with common metal choices being TiN for pFETs and Al-based alloys for nFETs. A common TiN layer can fulfill the dual role of workfunction metal in pFETs and barrier layer in nFETs, since the use of a thin (few nm) TiN layer allows modulation of the workfunction by deposition of an additional metal layer, especially after annealing [14,15]. In general, solutions that use one common workfunction metal and adjust the EWF of nFETs or pFETs by subsequent treatments (e.g. depositions or implantations) are preferred because they avoid damage to the high-k dielectric by the metal etch.

Two variants of the gate-last process have been developed, differing in the time of the high-k deposition. In the high-k first variant, which has been the first implementation of high-k technology to go in production [1], the high-k dielectric is deposited before the dummy poly-Si gate and is kept during the gate replacement. In the high-k last variant, introduced at the 32 nm node by Intel [11], the high-k dielectric is deposited after removal of the poly-Si gate, with the advantage of reducing the thermal budget to which the high-k material is exposed.

The gate-last integration scheme requires a substantial process development effort, given the additional process complexity, but allows minimization of the temperature to which the high-k/metal stack is exposed. This is highly beneficial to the resulting device properties, especially in setting the proper pFET workfunction at low EOT (see Section 1.2.3). The downside of the reduced thermal budget is that the quality of the Si/SiO$_2$ interface is not improved by the high-temperature densification as in the gate-first process, but proper annealing conditions have proven successful in improving interface state density and mobility [8]. Capping-layer-induced flatband voltage shift is also not achievable at typical gate-last thermal budgets [16–18], but the low budget in turn renders a dual-metal process effective in obtaining symmetric threshold voltages, although this is increasingly challenging for scaled and three-dimensional devices [19–21]. Scavenging techniques compatible with low-temperature processing have also been developed (see Section 1.3). From the point of view of the gate electrode, the gate-last scheme combines the advantages of a poly-Si electrode in the FEOL process with those of low-resistivity gate metal interconnect. Finally, the RMG process has been shown to be a useful technique for inducing channel strain in scaled devices [22,23].

1.2 Effect of high-k dielectrics on device properties

The introduction of high-k/metal gate technology has a profound effect on many critical device properties, especially equivalent oxide thickness (EOT), channel mobility, threshold voltage and reliability. As a matter of fact, high-k/metal gate technology is typically implemented in the form of a complex multilayer stack, where the desired overall electrical properties are obtained by designing and controlling the structure and properties of the interfacial layer, the bulk high-k layer, the optional dielectric cap and the gate electrode.
1. State of the art in high-k/metal gate technology

1.2.1 Equivalent oxide thickness

The necessity to compare dielectrics having different dielectric constants has led to the development of figures of merit that express the suitability of the dielectrics from the point of view of $C_{ox}$ scaling. The most commonly used one is EOT, which is defined as the equivalent SiO$_2$ thickness which would provide the same gate capacitance as the actual high-k dielectric:

$$EOT = \frac{k_{SiO_2}}{k_{HK}} t_{HK}$$

(1.1)

where $k_{SiO_2}$ and $k_{HK}$ are the relative dielectric constants of SiO$_2$ and of the high-k oxide, and $t_{HK}$ is the physical thickness of the high-k oxide. Capacitance equivalent thickness (CET) is also commonly used as a figure of merit of the gate stack, since it reflects the total capacitance that couples the gate to the channel charge. CET is defined as the equivalent SiO$_2$ thickness which would provide the same gate capacitance as the entire gate stack:

$$CET = \frac{k_{SiO_2}}{C_{inv}}$$

(1.2)

where $C_{inv}$ is the gate-to-channel capacitance density in inversion. CET is a measure of the electrical thickness in inversion (therefore another commonly used symbol for CET is $T_{inv}$), which differs from EOT by the additional contribution of the vertical charge distribution in the channel and of the depletion region in the gate electrode (if poly-Si is used as gate electrode). CET can be directly calculated from the measured capacitance, whereas extraction of EOT from electrical measurements generally requires the use of a quantum-mechanical simulator.

The reason behind the introduction of high-k dielectrics in CMOS logic was the capability to support EOT scaling below $\sim 1.2$ nm while keeping the physical thickness (and thus the tunneling-dominated leakage current) at acceptable levels. However, direct deposition of HfO$_2$ on Si was found to provide poor quality of the interface with Si, resulting in degraded interface state density. A high density of interface states is unacceptable in MOS transistors since it leads to increased Coulomb scattering (thereby degrading channel mobility, see Section 1.2.2) and to an unwanted additional conduction path in the off state (thereby degrading subthreshold slope and $I_{OFF}$). Although Hf silicates can provide better interface quality at the expense of lower dielectric constant, deposition of any Hf-based dielectric directly on Si was seen to cause undesired effects such as phase separation in the oxide and uncontrolled Si oxidation upon thermal annealing [24]. Therefore, current implementations of Hf-based high-k/metal gate technology rely on a thin SiO(N) interfacial layer (IL), which sets the quality of the interface with Si and allows a high-quality deposition of HfO$_2$ by ALD [25]. The physical thickness of the SiO(N) interfacial layer is object of a compromise between EOT scaling and other critical device properties, such as channel mobility, threshold voltage control and reliability (see Sections 1.3-1.4).
1.2. Effect of high-k dielectrics on device properties

1.2.2 Channel mobility

Relevance as a metric

The first rationale for gate length scaling in MOSFETs comes from the dependence of drain current on gate length, here shown in the saturation region for a conventional long-channel device model:

\[ I_D = \frac{1}{2m} \mu C_{ox} \frac{W_G}{L_G} (V_G - V_T)^2 \]  

(1.3)

where \( I_D \) is the drain current, \( m \) is the body-effect coefficient, \( \mu \) is the channel mobility, \( C_{ox} \) is the gate capacitance density, \( W_G \) is the gate width, \( L_G \) is the gate length, \( V_G \) is the applied gate bias and \( V_T \) is the threshold voltage. As shown in (1.3), drain current is inversely proportional to gate length and directly proportional to channel mobility, therefore reducing \( L_G \) and/or increasing \( \mu \) are effective ways to increase drive current and therefore circuit speed.

However, (1.3) is only valid for long-channel MOSFETs, where the lateral electric field is negligible compared to the vertical field. In scaled devices, the gate length becomes comparable to the mean free path of carriers in the channel, leading to off-equilibrium transport and eventually ballistic transport [26]. Ballistic theory predicts that drain current should be independent of both \( L_G \) and \( \mu \) in nanoscale MOSFETs, however Monte-Carlo simulations have shown that scattering can not be neglected even at \( L_G = 14 \) nm [27] and manufactured short-gatelength devices have been shown to operate in a quasi-ballistic regime, where gate length and scattering are still relevant [28–30].

Therefore, gate length scaling is still beneficial to device and circuit performance, as is a high effective channel mobility. As a matter of fact, mobility enhancement techniques have been introduced since the 90 nm node in order to improve device performance, in the form of strained Si channels obtained through different technological solutions such stress liners and regrown source/drain regions [31,32]. High mobility materials (such as Ge and III-V’s) are also considered for introduction in future technology nodes to obtain higher device and circuit performance than is possible on Si.

Impact of high-k dielectrics

Channel mobility in conventional SiON/poly-Si MOSFETs is mainly limited by three scattering mechanisms: Coulomb scattering by charged impurities in the substrate, which is critical at low effective field; phonon scattering by lattice vibrations in the substrate, which is dominant at intermediate effective field; surface roughness scattering from imperfections of the Si/SiON interface, which limits mobility at high effective field. The effective field \( E_{eff} \) is the average vertical electric field experienced by carriers in the channel, defined as [33]:

\[ E_{eff} = \frac{Q_B + \eta Q_{inv}}{k_{Si}\epsilon_0} \]
1. State of the art in high-k/metal gate technology

![Figure 1.2. Factors affecting channel mobility in high-k/metal gate stacks.](image)

In addition to scattering from surface roughness and interface traps, also present in conventional SiO$_2$-based gate stacks, additional scattering sources have been identified: low-frequency optical phonons (modeled as RPS); high bulk and interface charge and dipole densities (modeled as RCS); and local EOT variations caused by surface roughness at the IL/high-k and high-k/metal interfaces and by fluctuations of the dielectric constant (modeled as RSRS). Reproduced from [34] (University of Glasgow).

where $Q_B$ is the depletion charge density, $Q_{inv}$ is the inversion charge density, $k_{Si}$ is the relative dielectric constant of Si, $\epsilon_0$ is the vacuum permittivity, and $\eta$ is an averaging factor, usually taken as $1/2$ for electrons and $1/3$ for holes. In scaled devices, the depletion charge density is typically negligible compared to the inversion charge density, especially at high values of the effective field, which are the most relevant from the point of view of device operation. This, and the need to compare technologies where the standard split-CV method can not be used to calculate the depletion charge (e.g. SOI), has led to the modern tendency to just use $N_{inv} = Q_{inv}/q$ as measure of the effective field. Therefore, point values of the effective field which were classically used for comparison (e.g. 0.5 MV/cm for low fields and 1 MV/cm for high field) have been equated to corresponding point values of inversion charge density (e.g. $3 \times 10^{12}$ cm$^{-2}$ for low field and $1 \times 10^{13}$ cm$^{-2}$ for high field).

The introduction of high-k/metal gate technology has had a strong impact on carrier mobility, since the quality of the deposited high-k dielectric is typically worse than that of thermally grown SiO$_2$, giving rise to degraded interface state density and surface roughness and consequently lower mobility. Additionally, new sources of channel scattering have been identified in high-k/metal gate stacks (Fig. 1.2), motivating the definition of three new scattering mechanisms: remote Coulomb scattering (RCS), remote phonon scattering (RPS) and remote surface roughness scattering (RSRS).
Remote Coulomb scattering  RCS is the effect exerted by charges in the gate stack on the movement of carriers in the channel. The interaction of the carriers with the electric field created by the charges causes an increased scattering rate, which has been observed as the dominating mechanism limiting mobility at low effective field in high-k/metal gate stacks [35,36]. RCS is usually attributed to the presence of both fixed charges and interfacial dipoles, since the density of fixed charges that would explain the experimentally measured mobility is too high compared to the results of charge trapping measurement and to observed flatband voltage shifts [37]. RCS has also been observed in SiON/poly-Si stacks with ultra-thin gate oxides due to depletion charges in the poly-Si layer [38], but its entity is higher in high-k/metal gate stacks due to the intrinsically higher defect density in high-k dielectrics. As a matter of fact, high-k dielectrics suitable for CMOS applications derive their higher dielectric constant from the polarizability of ionic M-O bonds, which on the other hand provide less capability for relaxation and defect annihilation than SiO$_2$ [9]. RCS can be addressed to a certain extent by means of process optimization, i.e. by finding optimum deposition and annealing conditions that minimize the defect density in the oxide. Similarly to Coulomb scattering from charged impurities in the substrate, remote Coulomb scattering rates decrease with increasing effective field, thanks to the screening effect at high inversion charge density.

Remote phonon scattering  RPS is an intrinsic scattering mechanism in high-k devices, in the sense that it can not be avoided by means of process optimization. High-k dielectrics are characterized by ionic M-O bonds, since the ionicity and polarizability of the atomic bonds is the main cause for increased dielectric constant at the frequencies of interest for CMOS logic applications [39]. However, the contribution of polarizable ionic bonds to the permittivity is linked to low-frequency or soft optical phonon modes, whose frequency is much lower than the corresponding modes in SiO$_2$ and allows a strong interaction with carriers in the channel, effectively increasing their scattering rate. RPS was initially thought to be a strong limitation of high-k dielectrics because coupling of phonon modes in the dielectric with plasmon modes in the poly-Si gate led to unacceptably reduced channel mobility [40]. However, the introduction of metal gates (which possess much higher plasmon frequency) substantially mitigated the issue, although RPS still limits mobility in high-k/metal gate MOSFETs at intermediate effective field [41].

Remote surface roughness scattering  RSRS is caused by roughness at the IL/high-k and high-k/metal interfaces, which leads to local variations of the surface potential and consequently increased scattering rates. RSRS is dominant at high effective field, when carriers travel closer to the surface, while it is negligible compared to the other mechanisms at low effective field. The RSRS model can also account for local variations of the dielectric constant in the oxide (e.g. because of phase separation), since they can be seen as local variations of the electrical thickness of the oxide [35].
1. State of the art in high-k/metal gate technology

Figure 1.3. Schematic effect of remote scattering mechanisms on channel mobility. The additional scattering mechanisms cause reduced mobility over the whole effective field range in high-k/metal gate stacks compared to conventional SiO$_2$/poly-Si stacks. Reproduced from [13] (IBM).

Fig. 1.3 depicts the net effect of the integration of high-k dielectrics on carrier mobility, causing in general lowered mobility over the whole effective field range, although proper design of the gate stack and careful process optimization can lead to acceptable mobility for device operation. The most important design choice is the distribution of the EOT budget between the interfacial layer and the high-k layer. As a matter of fact, the nature and thickness of the high-k layer obviously have a strong impact on all three remote mechanisms [25, 42], with thin layers being preferrable because of the reduced number of scattering sites. However, a proper design of the interfacial layer can be even more effective on channel mobility. SiO(N) is commonly used as interfacial layer, which means that the interfacial layer itself does not cause additional scattering compared to the traditional SiON gate dielectric, therefore its net effect is to place the scattering sources in the high-k dielectric further away from the channel. A strong dependence of channel mobility on the interfacial layer thickness has been observed experimentally [43, 44] and has been explained theoretically as an exponential decay of the scattering potential over the SiO$_x$ thickness [38, 40].

1.2.3 Threshold voltage

High-k dielectrics were initially planned for integration with poly-Si gate electrodes, while metal gates were expected to be introduced at a later technology node to address the issues of gate depletion, boron penetration and high sheet resistance [45]. However, strong pinning of the Fermi level at the high-k/poly-Si interface was observed, leading to strong difficulties in properly setting the threshold voltage.
1.2. Effect of high-k dielectrics on device properties

This, in addition to the instability of the high-k/poly-Si system at high temperature, led to the cointegration of metal gates with high-k dielectrics [9].

Even when metal gates are used, however, shifts are observed in the flatband voltage compared to what one would expect from the vacuum workfunction of the metal. This is especially visible in gate-first processing, where, in addition to Fermi level pinning, flatband voltage roll-off has been observed for EOT < 3 nm, which has been attributed to oxygen vacancy generation in the SiO$_x$/HfO$_2$ dielectric stack and formation of dipoles at the IL/high-k interface [46–48]. The contribution of these mechanisms to flatband voltage is conceptually different from the effect of fixed charges and is included in classical flatband voltage models by introducing an effective work function (EWF) which takes into account the vacuum workfunction of the metal, Fermi-level pinning at the metal/high-k interface and the contribution of dipoles [49].

The roll-off phenomenon renders band-edge workfunctions for nFETs and pFETs difficult to achieve in gate-first integration schemes. A solution has been found in the integration of dielectric capping layers, since the high-temperature dopant activation anneal causes metal atoms from the capping layers to diffuse through the HfO$_2$ layer and reach the interfacial layer. The diffusion induces the formation of interfacial dipoles that can effectively shift the flatband voltage in either direction, depending on the choice of the dielectric cap [17, 43]. There is, however, a strong difference between capping layer technology for nFETs and pFETs: suitable materials for nFET capping layers include reactive oxides, such as La$_2$O$_3$ and MgO, while pFET dielectrics, such as Al$_2$O$_3$, are much less reactive. Such a difference is readily visible in the temperature requirements of the diffusion process, which can be fairly low for nFETs, while annealing at 1000 °C for a few seconds is necessary for pFETs. More importantly from the point of view of device performance, nFET capping layers react with the interfacial layer increasing its dielectric constant, with the net effect of shifting the threshold voltage while keeping EOT constant or even slightly decreased. On the other hand, pFET capping layers do not readily react with the interfacial layer and result in increased EOT, so that only a limited flatband voltage shift can be obtained in practice [50]. Additionally, integration of Al$_2$O$_3$ capping layers has been linked to mobility degradations due to increased remote Coulomb scattering [43]. Therefore, flatband voltage roll-off for pFETs is still problematic for gate-first integration at EOT lower than ~3 nm, although the issue can be mitigated at the expense of additional processing for pFETs, i.e. by dual channel integration [51, 52].

Gate-last processes are usually considered to be immune to vacuum workfunction degradation thanks to the low thermal budget after deposition of the gate metal, so that cap-free dual-metal processes are the most common implementation of gate-last high-k/metal gate technology [1]. However, flatband voltage roll-off in pFETs has been recently observed also in gate-last technology, although at much lower EOT (~0.8 nm). The same mechanism of oxygen vacancy diffusion and dipole formation as in the gate-first process is believed to be responsible for degradation of the gate-last EWF [53].


1. State of the art in high-k/metal gate technology

1.2.4 Reliability

The implementation of a novel material in CMOS technology is subject to the requirement of reliable long-term device operation, since integrated circuits are expected to have an operative lifetime of 10 years. Although several conditions determine the overall reliability, the gate stack has a direct influence on two critical aspects, i.e. oxide integrity and threshold voltage instability. The introduction of high-k/metal gate technology in manufacturing has thus required the evaluation of its impact on the aforementioned properties, quantified by means of time-dependent dielectric breakdown (TDDB) and bias-temperature instability (BTI).

High-k dielectrics usually possess intrinsically lower oxide integrity than SiO$_2$, observed as a lower breakdown voltage or time to breakdown. However, the integration of SiO(N) interfacial layers is effective in obtaining acceptable TDDB lifetimes [54].

BTI in pFETs, commonly referred to as negative bias-temperature instability (NBTI) due to the stress conditions, is remarkably similar in SiO$_x$/HfO$_2$/TiN and SiON/poly-Si stacks, which is compatible with the commonly accepted interpretation of instability and defect generation at the Si/SiO$_2$ interface being the main cause of unstable pFET threshold voltage. On the other hand, BTI in nFETs, commonly referred to as positive bias-temperature instability (PBTI) due to the stress conditions, is strongly degraded in high-k/metal gate MOSFETs. While PBTI is not a reliability concern in conventional gate stacks employing SiO$_2$ or SiON as gate dielectric, stable threshold voltages are challenging to achieve in SiO$_x$/HfO$_2$ nFETs under PBTI stress conditions. The effect has been attributed to a high density of bulk traps (mainly identified as oxygen vacancies) in HfO$_2$, which causes enhanced electron trapping and consequent threshold voltage shift under PBTI stress. The presence of a sufficiently thick interfacial layer is extremely important in achieving acceptable PBTI lifetime, since it regulates the tunneling probability from the channel to the high-k bulk traps [55].

1.3 Scalability issues in SiO$_x$/HfO$_2$ stacks

The previous section dealt with the various effects that the integration of high-k/metal gate stacks has on device properties and the design solutions which are used in current Hf-based technology to achieve the desired performance targets. This section presents an overview of the scaling issues which have been identified in the widespread SiO$_x$/HfO$_2$ dielectric stack and which can limit its applicability to future technology nodes.

Continuous EOT scaling is a key requirement on the International Technology Roadmap for Semiconductors (ITRS), being a crucial parameter in limiting both short-channel effects and gate leakage current. Even though the introduction of novel device architectures (e.g. multiple gate field-effect transistors (MuGFETs) and silicon on insulator (SOI)) has partly relaxed the EOT requirements, values lower than 0.6 nm are expected in 2020 according to the 2013 Edition of the ITRS [6].
Since the physical thickness of the HfO$_2$ layer has already reached its practical limits (1.5–2 nm due to considerations on direct tunneling, uniformity and reproducibility [56]) in current technology, EOT scaling is achieved by reducing the physical thickness of the SiO$_x$ interfacial layer. Nitridation can also be used to reduce the equivalent thickness of the interfacial layer, but the benefit is balanced by the introduction of additional N scattering centers and consequent mobility degradation [8].

### 1.3.1 Scaling of the interfacial layer

The SiO$_2$ interfacial layer is usually grown by low-temperature chemical oxidation, which has been shown to provide higher quality at lower physical thickness compared to rapid thermal oxidation or etched-back thermal oxide [57,58]. Several chemical oxidation techniques have been developed to achieve a thickness target between 0.4 and 1 nm, with the most commonly used being the so-called imec clean, consisting in a low concentration of O$_3$ diffused in a room-temperature DI-H$_2$O bath [59].

An alternative method for chemical oxidation was developed in this work (Paper I). The method consists in exposing the Si substrate to a mild oxidizing ambient containing O$_3$, O$_2$ and H$_2$O at 350°C in the ALD reactor. The temperature was chosen equal to the HfO$_2$ deposition temperature, rendering the technique suitable for in situ oxidation of the substrate before HfO$_2$ deposition, with the advantage of avoiding exposure of the interface to contaminations.

In gate-first processes, the chemical oxidation target is set to $\sim$1 nm in order to improve its uniformity and the quality of the HfO$_2$ deposition [60], but the final interfacial layer thickness is reduced by means of scavenging. This process consists in the integration of an oxygen scavenging material in the gate stack, so that the subsequent annealing step causes diffusion of oxygen from the interfacial layer to the scavenging element, thereby reducing the interfacial layer thickness. Scavenging can be obtained by integrating several elements in different positions of the gate stack, but the best results have been obtained by remote scavenging (e.g. by deposition of a thin Al layer on top of a thin TiN gate electrode) which can achieve well-controlled interfacial layer thickness without causing additional device degradation [13,43] (Fig. 1.4). While scavenging was initially developed for gate-first processes, the ongoing shift of a good part of the industry to gate-last integration schemes has motivated the development of low-temperature scavenging processes [53,61,62].

Hf-based stacks having no interfacial layer have also been demonstrated [63,64]. The result is achieved through accurate control of the deposition, the metal stack and the annealing environment and leads to EOT as low as 0.5 nm. However, it is very challenging to guarantee adequate channel mobility, threshold voltages and device reliability [13].

In general, reduction of the interfacial layer thickness has been shown to cause degradations in mobility, threshold voltage control and reliability, posing serious concerns about the long-term scalability of the SiO$_x$/HfO$_2$ dielectric stack.
1. State of the art in high-k/metal gate technology

**Figure 1.4.** Proposed model of interfacial layer scaling by remote scavenging. Reduction of SiO\(_x\) results from transfer of oxygen from the interfacial layer to the scavenging element via oxygen vacancies in HfO\(_2\). Reproduced from [43] (IBM).

**Effect on channel mobility** The interfacial layer has the beneficial effect of increasing the distance between the remote scattering centers in the high-k layer and the channel [40,65], thereby reducing the effect of the remote scattering mechanisms outlined in Section 1.2.2 on channel mobility. The measured mobility degradations at low EOT have been observed to follow a universal trend (Fig. 1.5), where the reduction in e.g. electron mobility has been quantified as 20 cm\(^2\)/Vs per 0.1 nm reduction in interfacial layer thickness [13]. The mobility reduction with decreasing EOT can be uniquely linked to decreasing interfacial layer thickness, since reduction of the high-k thickness has been shown to have the opposite effect on mobility, as it results in decreased number of scattering centers [25]. The observed mobility degradation is a concern for future scalability of the SiO\(_x\)/HfO\(_2\) stack, but computations of the actual effect of long-channel mobility on short-channel drive current have shown that the reduction is still acceptable given the advantage of a reduced EOT [30].

**Effect on threshold voltage control** The interfacial layer has been found to play a critical role in setting the threshold voltage. It is not only necessary for the formation of dipoles from integration of capping layers, but its thickness is also the main factor limiting flatband voltage roll-off in both gate-first and gate-last integration schemes. Proper annealing steps can be used to improve the quality of the interfacial layer and reduce the roll-off behavior, but aggressive scaling of the SiO\(_x\) IL below 0.4 nm will probably render flatband voltage roll-off unavoidable even in a gate-last process [53]. Threshold voltage control may then limit the scalability of the SiO\(_x\)/HfO\(_2\) dielectric stack, although the integration of dual-channel technology has been proposed as a way to overcome the issue [51].
1.4. Research directions

Figure 1.5. EOT-mobility trends for scaled gate stacks employing scavenging and La$_2$O$_3$ caps. Scaling of SiO$_x$/HfO$_2$ stacks can be achieved by aggressive scaling of the interfacial layer at the cost of strongly degraded channel mobility. The observed mobility trend is, however, still compatible with the minimum requirements for current technology nodes. Adapted from [13] (IBM).

**Effect on device reliability**  The degradation of device reliability with EOT scaling has been strongly linked to the thickness and quality of the interfacial layer [54]. The interfacial layer is critical in setting the quality of the interface toward Si (which is the main factor determining NBTI) and in limiting the gate leakage current (which in turn affects both TDBB and PBTI lifetimes). Reliability has been identified as the main problem limiting the scalability of the SiO$_x$/HfO$_2$ stack, since BTI lifetimes are predicted to degrade by 50-100x per 0.1 nm reduction in interfacial layer thickness (Fig. 1.6).

1.4 Research directions

A typical SiO$_x$/HfO$_2$ dielectric stack with an EOT of ~0.8 nm, designed according to the principles discussed in the previous section, employs a ~0.4 nm thick SiO(N) interfacial layer, leading to the consideration that, in the current Hf-based technology, the interfacial layer and the bulk high-k layer account for ~50% of the total EOT each. This consideration has motivated recent research to fork in two directions: on one hand, substitution of HfO$_2$ with a higher-k dielectric; on the other hand, substitution of the SiO(N) interfacial layer with a high-k IL.
1. State of the art in high-k/metal gate technology

Figure 1.6. Predicted change in BTI and TDDB lifetimes caused by variation of the SiO\textsubscript{x} interfacial layer thickness, evidencing the strong exponential dependence on IL thickness. Device reliability may set the ultimate limit to the scalability of the SiO\textsubscript{x}/HfO\textsubscript{2} dielectric stack. Reproduced from [54] (IBM).

1.4.1 Higher-k dielectrics

Early research on high-k oxides outlined the requirements for candidate gate dielectrics in CMOS technology [66–69]. Several characteristics have to be considered, regarding both intrinsic material properties and extrinsic processing and device considerations, with often contradictory indications. The most pressing requirements can be classified as follows.

Dielectric constant, band gap and band offsets The observed relationship between dielectric constant and energy gap has prompted the conclusion that high-k oxides with \( k \approx 20 – 30 \) provide the best balance between EOT scaling and leakage current control [66].

Thermodynamic stability Theoretical and experimental observations have restricted the list of reasonable candidates to those that are thermodynamically stable in contact with Si [70,71]. Even when only stable materials are considered, careful process development is necessary to avoid undesired interfacial reactions upon deposition or annealing.

Phase, lattice mismatch, oxygen permeability These are interrelated properties which depend both on the oxide and the deposition method. Research has mostly focused on amorphous materials, although polycrystalline dielectrics are
also acceptable, unless the grain boundaries result in unwanted leakage and oxygen diffusion paths. Epitaxial dielectrics can also be used with the purpose of obtaining higher dielectric constant, but can lead to problems such as defect creation by lattice mismatch, thickness dependence of the dielectric constant, and availability of deposition methods with the required uniformity, reproducibility and high throughput.

**Bulk and interface defect density**  The oxide should provide low defect density both in the bulk and at the interfaces. Although high-k oxides are by their own nature more defective than SiO$_2$, reasonably low bulk defect density can be achieved by optimization of the deposition and annealing conditions. Direct deposition of high-k dielectrics on Si usually results in unacceptably high interface state density, leading to the conclusion that integration of higher-k oxides will probably require the presence of an interfacial layer.

**Process compatibility and effect on device properties**  Candidate oxides need to be compatible with standard CMOS process requirements, such as the existence of suitable etching strategies. Finally, their impact on device properties (mainly EOT, mobility, threshold voltage and reliability) has to be fully assessed.

The largest number of studies on higher-k oxides can be divided in two groups: on one hand, many research groups are working on evaluating completely novel oxides, typically binary or ternary compounds of La [72–74]; on the other hand, it is possible to increase the dielectric constant of HfO$_2$ by phase engineering, i.e. by inducing crystallization in the higher-k tetragonal or cubic phase [75–78]. The first approach is less constrained by current technology and could potentially lead to superior solutions, but it also bears the burden of requiring the development of a completely new technology, especially for threshold voltage control. The second approach has the advantage of preserving compatibility with existing technology, but presents challenging scalability in the sub-nm EOT regime, due to thickness-dependent crystallization behavior.

### 1.4.2 High-k interfacial layers

The other approach to improving the scalability of the SiO$_2$/HfO$_2$ stack consists in replacing the SiO(N) interfacial layer with a high-k dielectric. This has the potential advantage of retaining full compatibility with current Hf-based technology, provided that the impact of the novel interfacial layer on device parameters is fully qualified.

While the same general requirements for high-k dielectrics outlined in the previous section are in principle valid for interfacial layers, their relative importance is quite different. As a matter of fact, a dielectric constant $\sim$10 is acceptable for the interfacial layer, since it would already be competitive toward SiO(N), while it is more important that the layer provide a high-quality interface toward Si.
1. State of the art in high-k/metal gate technology

This consideration is at the basis of the research on high-k interfacial layers: on one hand, some research groups are evaluating thermodynamically stable dielectrics, such as SrO and BeO, with the purpose of forming an atomically flat interface with Si [79, 80]; on the other hand, materials which have non-zero but limited reactivity with Si could be employed to form silicates [43, 81]. Silicates can potentially be an optimum choice for a high-k interfacial layer, since they provide reasonably high dielectric constants while having similar properties to SiO$_2$, especially in the capability to form a high-quality interface with Si and in the lack of low-frequency phonon modes that can reduce channel mobility [40]. These observations lead to the promise that a silicate interfacial layer could provide electrical properties comparable to a thicker SiO$_2$ layer but at lower EOT. The present work fits in this research branch, since the main result has been the identification, characterization and assessment of thulium silicate as a high-k interfacial layer.
Chapter 2

Integration of TmSiO as high-k interfacial layer

In Chapter 1, the current state of high-k/metal gate technology for CMOS logic applications has been discussed, focusing on the widespread SiO$_x$/HfO$_2$ dielectric stack. Possible solutions to gate stack design for future technology nodes have also been presented, among which the integration of high-k interfacial layers has promising potential.

This chapter discusses the design process and materials considerations for the choice of a novel high-k interfacial layer, followed by the development and evaluation of a CMOS-compatible process flow for integration of thulium silicate interfacial layers. Finally, compatibility with standard gate-last and gate-first integration schemes is addressed.

2.1 Design guidelines for a high-k interfacial layer

The integration of high-k interfacial layers has been proposed as a medium to long term solution to improve the scalability of high-k/metal gate stacks. The long-term applicability depends on whether novel higher-k materials will be successfully integrated, otherwise zero-IL HfO$_2$-based stacks may be necessary at a certain point, although at the expense of problematic reliability and threshold voltage control.

2.1.1 Design considerations

Some useful design considerations can be derived by making a qualitative calculation of the expected minimum EOT of different stacks. A scaled SiO$_x$/HfO$_2$ stack has a ~0.4 nm thick SiO$_x$ interfacial layer and a 1.5–2 nm HfO$_2$ layer. The physical thickness of the latter can not be scaled much further due to considerations on deposition uniformity and reproducibility and leakage current control, therefore the contribution of the HfO$_2$ layer to the total EOT can hardly be lower than...
2. Integration of TmSiO as high-k interfacial layer

~0.3 nm, assuming a dielectric constant of 18 which is reasonable for amorphous or monoclinic HfO$_2$. However, the 2013 Edition of the ITRS forecasts long-term EOT requirements as low as 0.41 nm and explicitly states that manufacturable solutions are not known that can satisfy the EOT target as early as 2024, when 0.49 nm EOT is expected [6]. In order to achieve very low EOT, three possible designs can be considered: aggressive use of scavenging, integration of a higher-k dielectric, and integration of a high-k interfacial layer.

**Scavenging** From the technological point of view, it is entirely possible to aggressively scale the interfacial layer to 0.2 nm or less by means of scavenging, thereby complying with the aforementioned ITRS requirement. This is the current industrial approach to EOT scaling and will likely be extended to the next technology nodes, but it does not constitute a long-term solution. As discussed in Chapter 1, scaling of the SiO(N) interfacial layer to less than 0.3–0.4 nm has been linked to strongly degraded channel mobility [13], threshold voltage control [53] and reliability [54], so that compliance with the required device performance targets is extremely problematic.

**Integration of higher-k bulk dielectrics** The integration of a higher-k dielectric would lower the contribution of the bulk high-k layer to the total EOT. However, if the same minimum thickness as the current HfO$_2$ layers has to be maintained, the EOT of the layer can only be reduced to 0.2 nm or 0.15 nm for a dielectric constant of 30 and 40 respectively. Since higher dielectric constants are not desired due to reduced band gap, a long-term stack design will still require integration of a high-k interfacial layer or the use of aggressive scavenging. The latter, however, is even more problematic in this case since higher-k dielectrics usually can not be placed close to the channel without strong transport degradations. As an example, remote phonon scattering rates have been shown to increase with increasing dielectric constant, further reducing channel mobility [40].

**Integration of high-k interfacial layers** High-k interfacial layers can be competitive toward conventional chemical oxides if they can provide an EOT contribution lower than 0.2 nm at a reasonably high physical thickness and with SiO$_2$-like properties in terms of interface state density, remote scattering and leakage current control. These desired properties point to the use of materials with a fairly low dielectric constant of 10-15 (since e.g. remote phonon scattering rates are negligible for low-dielectric-constant materials [40]), which in turn means that the physical thickness should be scalable and controllable in the sub-nm regime. Deposited layers such as SrO [79] and BeO [80] have been investigated and shown to provide good electrical properties, but it is extremely hard to deposit sub-nm films with excellent thickness control, uniformity, conformality and reproducibility. On the other hand, silicates can provide a manufacturable way to obtain well-controlled sub-nm interfacial layers, provided that a suitable oxide can be made to react with
Si in a highly controlled way. The resulting layer thickness would in this case be set by the reaction temperature, and precise thickness control could be obtained via proper annealing conditions.

2.1.2 Choice of material

As mentioned in the previous section, strong potential for extended scalability is held by the integration of silicates formed by reaction of a high-k oxide with the underlying silicon, as long as they can provide an interfacial layer which has comparable quality to chemical oxide but with higher dielectric constant. Although many oxides can react with Si, only some material classes can form silicates in a controllable way in the sub-nm regime. Among these, lanthanide oxides have been shown to form silicates which have good electrical quality and reasonably high dielectric constant [81].

A thorough comparison of silicates formed by the different lanthanide oxides, in terms of dielectric constant, band gap and offsets and interface quality, is still lacking in literature at the time of this writing, and a fair comparison would be especially difficult because of the strong dependence of the aforementioned properties on deposition and annealing conditions. However, it is possible to derive some guidelines for material selection from available data on lanthanide oxides [68,82,83] and on the basis of processing considerations.

The crucial requirement in the interfacial layer formation process consists in being able to form silicates of low thickness and high quality. To this aim, it must be considered that many lanthanide oxides possess the undesirable properties of hygroscopicity and (relatively) strong reactivity with Si. Hygroscopicity is undesirable because it can lead to degradation of the dielectric constant by hydroxylation and to reliability concerns; strong reactivity is undesirable because it renders thickness control in the sub-nm range difficult to achieve. Both properties tend to decrease with increasing atomic number across the lanthanide series, in a way that has been linked to the decreasing ionic radius [84–87]. For example, La$_2$O$_3$ presents the highest hygroscopicity and reactivity among lanthanide oxides, although it has been successfully employed to form La silicate (La$_x$Si$_y$O or simply LaSiO) interfacial layers, which present good electrical quality of the interface with Si and a reasonably high dielectric constant of 10-12 [81]. However, special in situ processing is necessary to avoid moisture absorption [88] and a strictly controlled composition of the metal stack and annealing environment is required to produce low and controlled silicate thickness [73,89]. Such processing complexities, and the lack of compatibility with pFET operation in Hf-based stacks [43], lead to the conclusion that LaSiO cannot be considered as a candidate for a general-purpose interfacial layer.

In this work, the selection of a candidate material has been driven by the considerations discussed above and has led to the choice of thulium silicate (Tm$_x$Si$_y$O or simply TmSiO) as a candidate high-k interfacial layer. Being the oxide of a high-atomic-number lanthanide, Tm$_2$O$_3$ is expected to possess low hygroscopicity and low reactivity with Si, which could enable its integration in a straightforward and
2. Integration of TmSiO as high-k interfacial layer

controllable process: indeed, the lowest reactivity with Si upon annealing has been reported for Tm$_2$O$_3$, within the class of binary lanthanide oxides [87]. The choice is also supported by the reasonably high dielectric constant of TmSiO (~12 [90], similar to LaSiO) and by the fairly high band gap and offsets of Tm$_2$O$_3$ (band gap of 6.5 eV, conduction band offset of 2.3 eV and valence band offset of 3.1 eV for MBE Tm$_2$O$_3$ on Si [91]; band gap of 5.3 eV, conduction band offset of 1.7 eV and valence band offset of 2.95 eV for ALD Tm$_2$O$_3$ on Ge [92]).

A common concern about the introduction of lanthanide elements in CMOS technology lies in their presumed scarcity, as indicated by the common name of “rare earths”. However, this name has a historical origin and relates to the geographical availability of mineral deposits rather than to the absolute abundance of the elements in the Earth’s crust. Even Tm, which is the rarest element among the lanthanide series, has comparable abundance to I and Hg and is ~200 times more abundant than Au. The tendency of lanthanide elements to be distributed across the crust rather than concentrated in economically exploitable ore deposits, together with geopolitical considerations on the distribution of the world’s resources and production, could pose a problem to the economic viability of Tm in the microelectronics industry. However, this is mainly a geopolitical consideration rather than a scientific and technological one and is part of the more general issue of the reliance of many high-tech industries on lanthanide elements.

Existing literature on CMOS applications of Tm-based oxides is extremely limited. The TmSiO/Tm$_2$O$_3$ dielectric stack has been integrated as gate dielectric [90], but the relatively low dielectric constant of Tm$_2$O$_3$ (16-18) strongly limits the scalability of the stack. Tm$_2$O$_3$ has also been investigated as capping layer for La-based gate stacks, where its integration provides an increased dielectric constant of the interfacial silicate layer and a corresponding decrease in EOT [93]. This work follows a completely different approach, since TmSiO is investigated for integration as a general-purpose interfacial layer, without requiring the use of Tm$_2$O$_3$ or La$_2$O$_3$ as bulk high-k layer, and with specific focus on applicability to industrially-relevant Hf-based gate stacks.

2.2 Integration of TmSiO in a CMOS process

The first requirement for a novel interfacial layer is the possibility to integrate it in a way that is compatible with the standard CMOS process flow. This translates, at a minimum, in the need for a suitable deposition technique and a way to selectively remove the material outside the gate area. However, in order to be competitive toward the conventional SiO$_x$ interfacial layer, the high-k interfacial layer should also introduce as little additional process complexity as possible.

To this aim, in this work, the thulium silicate process has been developed as a drop-in replacement for the conventional chemical oxidation, in the sense that the interfacial layer process is complete at the time of the deposition of the bulk high-k layer, without requiring special in situ processing and without posing any constraints
2.2. Integration of TmSiO in a CMOS process

Figure 2.1. Process flow for integration of TmSiO as interfacial layer in a generic high-k/metal gate stack. The conventional chemical oxidation is replaced by a three-step procedure consisting in deposition of Tm$_2$O$_3$, silicate formation by RTA and removal of excess Tm$_2$O$_3$. (Paper II)

on the composition of the rest of the gate stack. This kind of process design has the additional advantage of decoupling the choice of the two layers, making the TmSiO interfacial layer potentially compatible with future higher-k dielectrics (provided that the interaction between the two materials, especially upon annealing, is fully qualified).

The thulium silicate process has been developed as a sequence of 3 core steps (Fig. 2.1): Tm$_2$O$_3$ deposition, silicate formation and removal of excess Tm$_2$O$_3$. Additionally, surface cleaning before gate deposition and oxide removal outside the gate area are addressed.

2.2.1 Surface cleaning

Surface cleaning is crucial to the subsequent silicate formation, as it is necessary to ensure complete removal of native or chemical oxide. A standard pre-gate-oxidation cleaning can be used, followed by etching of chemical oxide in diluted HF chemistry to ensure deposition on H-terminated Si surface.

In this work, H$_2$SO$_4$:H$_2$O$_2$ (3:1 by volume) at > 100 °C was used for removal of organic contamination, while diluted (0.5-5%) HF was employed for removal of chemical oxide and particles [59]. Use of the more common RCA clean was also tested, performing the standard clean 1 (SC-1) in H$_2$O:NH$_4$OH:H$_2$O$_2$ 5:1:1 at 80 °C and the standard clean 2 (SC-2) in H$_2$O:HCl:H$_2$O$_2$ 6:1:1 at 80 °C, and including oxide removal (in 5% HF) and rinsing steps after each clean. Both cleaning procedures resulted in similar silicate formation, provided that a final step in diluted HF chemistry was always performed to ensure deposition on HF-last Si. Different
2. Integration of TmSiO as high-k interfacial layer

Concentrations of the final HF cleaning step can be used for improved process compatibility, as long as the chemical oxide formed during the cleaning procedure is completely removed. As a matter of fact, proper formation of thulium silicate could not be obtained on chemical oxide surfaces, at least at relatively low temperatures.

2.2.2 Deposition of Tm$_2$O$_3$

Although several deposition techniques are commonly employed in the literature to test novel materials, ALD is the technique of choice in the industry for deposition of high-k dielectrics, due to its excellent uniformity, conformality and reproducibility. Given the emphasis put in this work on applicability to actual CMOS process flows, ALD has been used to deposit Tm$_2$O$_3$. Since the only existing process was inadequate to the present application, a novel ALD process (described in Chapter 3) has been developed for deposition of Tm$_2$O$_3$ on HF-last Si.

2.2.3 Formation of thulium silicate

TmSiO has been formed by reaction of the Tm$_2$O$_3$ layer with the Si substrate (Fig. 2.2). Compared to direct deposition of a silicate layer, this process can provide higher interface quality and higher control on the resulting thickness in the sub-nm regime, by means of controlling the annealing temperature. The silicate formation was obtained by rapid thermal anneal (RTA) in inert environment (N$_2$), where the formation process presents the desirable properties of saturating time dependence and strong temperature dependence (Fig. 2.3). Such properties allow the use of reasonably long annealing times (e.g. 60 s) to improve uniformity and reproducibility, while the desired thickness is obtained by controlling the annealing temperature. The complete characterization of the TmSiO formation process is reported in Paper II and Paper IV, where it is also shown that the final TmSiO thickness is independent of the initial Tm$_2$O$_3$ thickness, provided that the oxide layer is not fully consumed by the reaction. This property is beneficial to the uniformity and reproducibility of the TmSiO process, since a selective etching solution is available to remove the excess Tm$_2$O$_3$.

Silicate growth in an oxidizing environment was also tested, but the lowest physical thickness was obtained by silicate formation in inert environment (Fig. 2.4). A similar process was also performed with La$_2$O$_3$ (since LaSiO interfacial layers have been investigated extensively in literature [43, 73, 81, 89]), showing that TmSiO can form thinner layers than LaSiO and achieve sub-nm physical thickness at 500–550 °C. An additional analysis has been performed to investigate the silicate formation mechanism. An exponential trend is observed for growth in inert environment, similar to what has been reported for LaSiO growth in forming gas anneal (FGA), where the Arrhenius dependence has been attributed to the temperature dependence of atomic diffusion in the silicate layer [81]. From the slope of the Arrhenius plot (Fig. 2.5), it is possible to extract values of activation energy of 0.31 eV and 0.25 eV for TmSiO and LaSiO respectively. The slightly higher activation energy for TmSiO
2.2. Integration of TmSiO in a CMOS process

**Figure 2.2.** TEM cross-section of TmSiO formed by reaction of Tm$_2$O$_3$ with the underlying Si. A graded silicate layer is formed, while no SiO$_x$ layer is visible. (Paper II)

**Figure 2.3.** Physical thickness versus annealing time for TmSiO layers formed by reaction of Tm$_2$O$_3$ with Si in inert environment at 550, 600 and 800 °C. The saturating time dependence can be helpful in achieving a uniform and reproducible process, while the desired thickness can be obtaining by setting the proper deposition temperature.
2. Integration of TmSiO as high-k interfacial layer

Figure 2.4. Physical thickness versus annealing temperature for TmSiO and LaSiO layers formed by reaction of Tm$_2$O$_3$ and La$_2$O$_3$ with Si in inert and oxidizing environment. Lower physical thickness is achieved by TmSiO formed in inert environment, where sub-nm physical thickness can be obtained at 500–550 °C.

Figure 2.5. Physical thickness of TmSiO and LaSiO versus the reciprocal of formation temperature. Silicate formation in inert environment obeys an Arrhenius-type temperature dependence, which is compatible with a growth model where Si diffusion in the silicate layer is the limiting mechanism.

is in line with its reported lower reactivity with Si, which has been explained by limited Si diffusion in the film [87].

2.2.4 Selective removal of Tm$_2$O$_3$

After TmSiO formation, it is necessary to remove the unreacted Tm$_2$O$_3$ so that only the desired TmSiO layer is left on the substrate. The etching solution has to be selective to both TmSiO and all other materials that are present on the wafer.
2.3 Compatibility with standard CMOS process flows

surface (where no TmSiO was formed since Si was not exposed). Although no specific studies on etching of Tm$_2$O$_3$ have been published, to the best of the author’s knowledge, a general conclusion from available literature on integration of lanthanide oxides is that dry etch can not be applied, due to the lack of volatile byproducts, while high etch rates can be obtained in low-pH solutions such as concentrated HCl and H$_2$SO$_4$ [94–96].

In this work, concentrated H$_2$SO$_4$ was used to remove the excess Tm$_2$O$_3$. This choice provides both high selectivity toward TmSiO (experimentally determined to be > 23:1) and full compatibility with standard CMOS processes, since in both gate-first and gate-last integration schemes only oxides and nitrides are present on the wafer surface at the time of the high-k deposition. Rutherford backscattering spectroscopy (RBS) measurements were performed after a prolonged etch in H$_2$SO$_4$, confirming that the TmSiO layer was not etched.

After removal of the excess Tm$_2$O$_3$, the interfacial layer module is complete and the bulk high-k layer and the rest of the gate stack can be deposited as prescribed by the process flow in use. No specific investigation of the effect of waiting time and exposure to air on device characteristics was performed, but from a practical point of view both should be kept to a minimum, since any scaled high-k/metal gate stack is to a certain degree sensitive to moisture absorption [97].

2.2.5 Removal of TmSiO

The last requirement to be addressed for integration of a novel material in the gate stack is the capability to pattern it. It is, in fact, necessary to remove the dielectric outside the gate area, since source and drain regions need to be accessed for silicidation in gate-first integration schemes, and contact areas need to be accessed in gate-last integration schemes (although CMP can be used to clear the dielectric outside the gate trench in “high-k last” process flows).

Etch tests performed on TmSiO showed that the silicate can be removed by a short dip in diluted HF chemistry, at a slightly lower etch rate than chemical oxide, thereby rendering the integration of the high-k interfacial layer straightforward from the point of view of patterning. As a matter of fact, the common dry-wet etching procedure for HfO$_2$, where the dielectric is damaged by plasma etching and the residuals are removed in diluted HF chemistry [94,98], was found effective in removing TmSiO as well.

2.3 Compatibility with standard CMOS process flows

As described in the previous section, the silicate formation module has been designed keeping compatibility to standard CMOS process flows as a priority, in order to develop a process for integration of a high-k interfacial layer that may be competitive toward the established chemical oxidation process.

Verification of the actual compatibility of the interfacial layer with standard process requirements, especially regarding the thermal budget, is the subject of
Paper II, where a simplified capacitor-only flow has been used in order to be able to fully characterize the process over the silicate formation and post metallization anneal (PMA) temperature spaces. The main results are reported here, while the reader is referred to the paper for the complete discussion.

### 2.3.1 Gate-last integration scheme

Compatibility with gate-last integration schemes has been verified by avoiding any thermal treatments after the gate stack formation, except for the final FGA at 400°C. Plasma-enhanced chemical vapor deposition (PECVD) SiO$_2$ was employed as bulk dielectric, in order to be able to qualify the intrinsic properties of TmSiO without undue influence from the properties of the bulk high-k dielectric. Parallel experiments were performed with thulium and lanthanum silicate, since LaSiO is commonly used in literature, leading to the conclusion that TmSiO is superior in terms of EOT (Fig. 2.6).

Thulium silicate formation at 500–550°C provided the best results, with sub-nm physical thickness, sub-0.3nm EOT and interface state density at flatband condition $< 2 \times 10^{11}$ cm$^{-2}$eV$^{-1}$. These preliminary results provided encouraging evidence of the potential benefits of integrating TmSiO in a gate-last MOSFET process, which will be discussed in the following chapters.

Direct calculation of silicate dielectric constant as ratio between IL EOT and physical thickness does not produce meaningful results, because of the inevitable inaccuracy of the physical thickness values obtained by spectroscopic ellipsometry, caused by measurement uncertainty at low thickness and model inaccuracy at high thickness. The latter is due to the fact that the models were calibrated on samples fabricated at low temperature, which are the most relevant from the point of view of EOT, while silicates formed at higher temperature are likely to present a different composition and a different optical response [99].

### 2.3.2 Gate-first integration scheme

Compatibility with gate-first integration schemes has been verified by performing a high-temperature PMA after deposition of a 100 nm thick TiN layer, simulating the effect of dopant activation anneal on a full metal gate stack. The results should also be applicable to MIPS stacks, since the poly-Si layer would replace the thick TiN layer as an oxygen diffusion barrier.

The experiment showed that thulium silicate interfacial layers are compatible with high-temperature annealing up to 1000°C, providing competitive values of EOT (Fig. 2.7) and interface state density (Fig. 2.8). Leakage current data (Fig. 2.9) show that the presence of a silicate interfacial layer is beneficial to gate current control over the whole PMA temperature range, although very high temperatures (1100°C) have a detrimental effect on gate current. The current data is useful in drawing a comparison among different interfacial layers and different annealing temperatures, but the very low values of leakage current density in these samples
2.3. Compatibility with standard CMOS process flows

Figure 2.6. Extracted EOT of the interfacial layer versus silicate formation temperature, for gate stacks integrating TmSiO and LaSiO as interfacial layers. TmSiO interfacial layers can achieve competitive EOT of $\sim 0.2$ nm when formed at 500–550 °C.

should not be taken as an estimate of the actual gate leakage current in MOSFETs, since the gate current in these samples is mainly limited by the presence of a 2 nm thick SiO$_2$ layer and by generation-recombination mechanisms in the substrate [100].
2. Integration of TmSiO as high-k interfacial layer

Figure 2.7. Extracted EOT of the interfacial layer versus PMA temperature, for gate stacks integrating TmSiO and LaSiO as interfacial layers. TmSiO contributes a lower EOT over the whole temperature range and achieves competitive values of 0.2–0.3 nm after annealing at 900–1100 °C. (Paper II)

Figure 2.8. Interface state density at flatband condition versus PMA temperature, for gate stacks integrating TmSiO and LaSiO as interfacial layers. All samples show $D_{it}$ of $\sim 1 \times 10^{11}$ cm$^{-2}$eV$^{-1}$, with a possible degradation after PMA at 1100 °C. (Paper II)
2.3. Compatibility with standard CMOS process flows

Figure 2.9. Leakage current density at 1 V in inversion versus PMA temperature, for gate stacks integrating TmSiO and LaSiO as interfacial layers, compared to a reference sample without interfacial layer and without annealing (“No IL”). The reduced leakage current in LaSiO/SiO$_2$ stacks compared to TmSiO/SiO$_2$ is caused by the higher physical thickness, and EOT, of the LaSiO layer. (Paper II)
Chapter 3

Atomic layer deposition of Tm$_2$O$_3$

Chapter 2 discussed the requirements for the TmSiO interfacial layer technology. One crucial requirement consists in the availability of a fab-friendly deposition technique for Tm$_2$O$_3$, i.e. one that is both technically suitable for the deposition of high-k gate dielectrics and economically viable for high-volume production. Several deposition methods are capable of depositing high-k dielectric materials with the required thickness control, e.g. Molecular Beam Epitaxy (MBE) and Pulsed Laser Deposition (PLD). However, atomic layer deposition (ALD) has become the main deposition technique for high-k dielectrics in the microelectronics industry thanks to its excellent uniformity and conformality, combined with the possibility to achieve sufficiently high throughput.

After a brief description of the basic principles of ALD, this chapter presents the ALD process development for deposition of Tm$_2$O$_3$. The film properties which are most relevant to CMOS processing are then analyzed, followed by a discussion on a possible way to improve the process.

3.1 Principles of atomic layer deposition

3.1.1 Basic technique and applications

Atomic layer deposition (ALD), also known as atomic layer epitaxy (ALE) and atomic layer chemical vapor deposition (ALCVD), was proposed in the late 1970’s mainly as a deposition technique for large-area displays [101]. More recently, ALD has found many applications in semiconductor processing, e.g. for deposition of high-k dielectrics in MOSFETs, copper diffusion barriers in BEOL interconnects and dielectrics in trench capacitors for dynamic random access memory (DRAM) [102]. The widespread use of ALD in such applications is driven by the high requirements in terms of uniformity and conformality, while the main drawback of the technique, i.e. its limited throughput, can be accepted since thick layers are not needed (for example, gate dielectrics are only few nm thick).
ALD is a variant of chemical vapor deposition (CVD). In ALD, the deposition is obtained via sequential cycles of self-limiting surface reactions, which are characterized by a saturating time dependence and possess an irreversible nature, at least on a limited time scale. These properties are responsible for the excellent uniformity, conformality and thickness control that can be achieved by ALD [103]. Self-limiting reactions are obtained by proper choice of the precursors and of the deposition temperature and lead to a well-determined increase in film thickness during each cycle (usually less than one monolayer, due to steric hindrance and availability of reactive sites).

Given the discrete nature of film thickness control and the strong dependence of the duration of each cycle on the specific design of the reactor, the deposition rate is typically expressed as thickness per cycle instead of thickness per time unit, with Å/cycle being the most common unit. Some of the literature on ALD defines the quantity growth per cycle (GPC) to avoid any ambiguity with the usual meaning of deposition rate, as defined in common CVD or physical vapor deposition (PVD) techniques. However, any (unlikely) confusion can be easily avoided by using the correct units of measure, therefore the more commonly used term deposition rate is employed in this work.

During each deposition cycle, the substrate is sequentially exposed to a series of reactive and inert gaseous species. For the common case of binary compounds, such as Tm$_2$O$_3$, the substrate is sequentially exposed to (Fig. 3.1):

1. First precursor, typically a metallic compound, which adsorbs on the substrate
2. Inert purging gas, which ensures complete removal of the first precursor and of possible byproducts from the reactor atmosphere
3. Second precursor, which reacts with the adsorbate to provide the desired compound (e.g. an oxidizing agent for oxide depositions)
4. Inert purging gas, which ensures complete removal of the second precursor and of byproducts from the reactor atmosphere.

The separation in time between the two precursors and the presence of the purge steps ensure that no reaction can take place in the gas phase, but only on the substrate. Besides the inherent lack of gas-phase reactions, another key difference of ALD compared to CVD is the temperature independence of the deposition rate, which is explained by the fact that each reaction is driven to completion, rendering the deposition rate essentially independent of both gas flow and surface reactivity. Such ideal behavior is only observed in the so-called ALD window, and a weak temperature dependence can still be observed in practice due to several non-idealities, such as temperature-dependent concentration of surface reactive sites, partial reversibility of the reactions or partial precursor decomposition [103]. Outside the ALD window, strong temperature dependence is observed, due to precursor condensation or lack of reactive sites at low temperature, and precursor decomposition or desorption at high temperature (Fig. 3.2).
3.1. Principles of atomic layer deposition

Figure 3.1. Illustration of ALD working principle. Film deposition proceeds by cycles of exposure to first precursor, inert gas purge, exposure to second precursor and inert gas purge. Reproduced from [104] (Pohang University of Science and Technology).

Figure 3.2. Temperature dependence of ALD deposition rate. The temperature range yielding a constant deposition rate defines the (a) ALD window, which is limited at low temperature by (b) precursor condensation or (c) insufficient surface reactivity, and at high temperature by (d) precursor decomposition or (e) desorption. Many actual ALD processes lack a well-defined ALD window and show a (f) continuous temperature dependence of the deposition rate, caused by a strong dependence of the surface reactions on the number of available reactive sites. Reproduced from [83] (Helsinki University of Technology).
Deposition of ternary oxides is usually performed by alternating one or more cycles of the oxide of the first element with one or more cycles of the oxide of the second element. The as-deposited nanolaminate structure can then be annealed in order to achieve better intermixing of the layers and improved stoichiometry (by means of oxygen diffusion into or out of the layer). The procedure is easily extended to more complex compounds and to precise doping of a host material with low amounts of a different element. Both applications are relevant to the research on high-k dielectrics, since higher-k oxides are often ternary compounds, while phase engineering of HfO$_2$ is achieved by doping with Al or other metals (see Section 1.4).

3.1.2 Technical considerations

As previously mentioned, the duration of each ALD cycle depends strongly on the design of the reactor, as does the achievable film uniformity. Several types of reactors have been designed to improve uniformity and throughput and are commercially available for semiconductor processing, with different gas distribution systems (cross-flow or showerhead), different heating systems (hot-wall or cold-wall reactors) and different processing modes (single-wafer or batch).

Reactor design in ALD is, in general, less critical than in CVD, since the reliance on self-saturating reactions means that neither gas flow uniformity nor temperature uniformity affect film uniformity as much as in a CVD reactor. However, non-idealities such as presence of dead pockets, improper gas mixing, temperature-dependent concentration of reactive sites, creation of reactive byproducts and precursor decomposition can lead to undesired loss of uniformity and point to the importance of a proper design of the reactor [105].

The most critical issue, affecting both cycle duration and film uniformity, is efficient removal of precursors and byproducts from the reactor atmosphere. ALD systems for CMOS applications typically operate at a pressure of $\sim$1 mbar, which provides a compromise between efficient use of N$_2$ as carrier gas and fast removal of reactants and byproducts from stagnant gas in the reactor [106]. Proper gas flow design can lead to shorter cycles without compromising uniformity, although the duration of each exposure and purge step greatly depends on the employed chemistry. As a matter of fact, cycle times can easily vary from 1 s to a few min, using the same reactor, depending on the deposited material and on the required conformality.

Many ALD systems also offer the possibility to use plasma as an additional source of energy [107]. Plasma-enhanced atomic layer deposition (PEALD), as opposed to thermal ALD, employs capacitively or inductively coupled plasma to facilitate reactions that would normally not occur or would require higher temperatures. An example is the deposition of elemental metals, where precursor reduction can be achieved at reasonably low temperatures by hydrogen radicals created from H$_2$ plasma. However, PEALD presents disadvantages in terms of uniformity and conformality, therefore deposition of gate dielectrics by thermal ALD is generally preferrable, especially for application to three-dimensional structures such as FinFETs.
and nanowires.

In this work, the commercially available Beneq TFS 200 ALD system was used. The tool employs a crossflow-type 200 mm hot-wall reactor enclosed in a cold-wall chamber, with the possibility to perform both thermal and plasma-enhanced ALD from room temperature to 500 °C. The system supports 4 liquid sources and 2 solid sources, which can be heated up to 300 °C. An external BMT 802N ozone generator allows the use of O$_3$ as oxidizing agent. Ar is used as carrier and purge gas, at a base pressure of 1 mbar.

### 3.2 Process design for ALD Tm$_2$O$_3$

A necessary condition for the potential integration of TmSiO in CMOS manufacturing is the existence of an ALD process for deposition of Tm$_2$O$_3$ which can achieve the required thickness control, uniformity and conformality, especially for an eventual integration in three-dimensional device structures.

Prior to this work, only one process was available for deposition of Tm$_2$O$_3$ by ALD, providing a deposition rate of 0.22 Å/cycle at 300 °C [108]. The process, however, employed Tm(thd)$_3$ as precursor, which requires the use of O$_3$ as oxidizing agent. The use of O$_3$ is not desirable in the deposition of scaled high-k dielectrics, since it leads to the growth of a thick (1–1.5 nm) SiO$_x$ layer at the interface [109–113]. In the present application, the insufficient reactivity of Tm$_2$O$_3$ with chemical oxide (see Section 2.2) rendered the existing O$_3$-based process not compatible with the final purpose of forming a thin TmSiO layer in direct contact with Si. This consideration led to the development of a novel ALD process, which is presented in Paper III together with a thorough characterization of the film properties. Here, the design principles behind the developed process are described in detail.

#### 3.2.1 Choice of precursor

Different classes of metallic precursors have been employed for ALD, both inorganic and organic [103,114,115]. Several requirements need to be satisfied for successful use of a precursor in an ALD process, among which sufficient volatility, sufficient thermal stability, strong reactivity and high purity [116].

For deposition of lanthanide oxides, the lack of volatile inorganic compounds restricts the choice to metalorganic compounds. Within this class, the coordination chemistry of lanthanide complexes dictates the use of sterically demanding or chelating ligands in order to achieve sufficient volatility. Although several compounds have been investigated, the only precursors that are broadly available for most or all of the lanthanide elements employ one of the following three ligands [83,117]:

- precursors employing thd ligands have been successfully employed in the deposition of lanthanide oxides, but their limited reactivity requires the use of strong oxidizing agents, such as O$_3$, which are undesirable in the deposition of scaled high-k/metal gate stacks, as previously discussed;
3. Atomic layer deposition of Tm$_2$O$_3$

Figure 3.3. TmCp$_3$ molecule and model of adsorption. Adsorption of the precursor to an available reactive surface site involves loss of one or more ligands by protonation, as supported by RGA measurements.

- precursors employing silylamide ligands achieve sufficient reactivity to allow the use of mild oxidizing agents, such as H$_2$O, but often result in high Si content in the deposited films;

- organometallic precursors based on the Cp ligand provide high reactivity, rendering them compatible with H$_2$O-based processes, and have been shown to provide low impurity content in the deposition of many rare earth oxides.

The second precursor in an ALD process is a gas-phase reactant which contains the non-metallic element of the desired compound. For the deposition of oxides, several oxidizing agents can be used as second reactant, among which the most common are H$_2$O, H$_2$O$_2$ and O$_3$ in thermal ALD and O$_2$ plasma in PEALD. H$_2$O is usually preferred in the deposition of gate dielectrics, since it minimizes or avoids the growth of SiO$_x$ at the interface, but stronger oxidizing agents may be required depending on the first precursor.

In this work, the choice of a Tm precursor for deposition of Tm$_2$O$_3$ was performed among commercially-available precursors. The only widely available Tm-containing precursors suitable for ALD were tris(2,2,6,6-tetramethyl-3,5-heptanedionato)thulium(III), tris[N,N-bis(trimethylsilyl)amide]thulium(III) and tris(cyclopentadienyl)thulium(III). On the basis of the previous discussion, the latter (chemical formula TmCp$_3$, Fig. 3.3) was chosen, since the organometallic compound presented the advantages of higher reactivity, compatibility with the use of H$_2$O as oxidant, and lack of potential impurities such as Si.

The manufacturer did not provide information on the thermal stability of TmCp$_3$,
3.2. Process design for ALD Tm$_2$O$_3$

Figure 3.4. Suitability of the main Cp-based precursors for deposition of lanthanide oxides. Compounds above the dashed line have sufficient thermal stability for use in ALD. Reproduced from [114] (Helsinki University of Technology).

but it was reasonable to expect sufficient stability for use as ALD precursor, given the available literature data on deposition of lanthanide oxides from organometallic precursors. Thermal stability of Cp-based rare earth compounds tends to degrade with increasing ionic radius, requiring the use of more sterically demanding ligands for deposition of e.g. La$_2$O$_3$ (Fig. 3.4). Sufficient thermal stability for TmCp$_3$ could thus be reasonably assumed from the reported stability of YCp$_3$ [118], since Tm$^{3+}$ has a slightly lower six-coordinate ionic radius than Y$^{3+}$ (102 and 104 pm respectively [119]) and the chemistry of Y is in many ways similar to that of high-atomic-number lanthanides [117].

It was later observed experimentally that the highly reactive TmCp$_3$ molecule also presents the strong advantage of enabling high-quality deposition on HF-last Si, with very short incubation time of 2-6 cycles. This facilitates uniform deposition of a thin Tm$_2$O$_3$ layer on H-terminated Si substrate, which is instead problematic in the deposition of other oxides, such as HfO$_2$, where a chemical oxide is necessary as starting surface to achieve a high-quality deposition [25,120]. The developed process is thus capable of depositing Tm$_2$O$_3$ in direct contact with Si, which enables its use in the formation of TmSiO interfacial layers.

3.2.2 Process characterization

The first step in the process development phase was the choice of the precursor source temperature, since TmCp$_3$ is in solid phase at room temperature and requires heating to achieve sufficient vapor pressure. Given the lack of information on vapor
3. Atomic layer deposition of Tm$_2$O$_3$

Figure 3.5. Deposition rate of Tm$_2$O$_3$ versus duration of the exposure to TmCp$_3$ and H$_2$O at 250 °C. Exposure time of 15–20 s and 4 s respectively were sufficient to guarantee self-saturating film growth in the ALD window 200–300 °C.

pressure from the manufacturer and the lack of a suitable source pressure sensor, the precursor temperature was determined experimentally, i.e. by finding the lowest temperature that would allow the precursor to adsorb on the whole substrate in a reasonable amount of time. Shorter exposure times could of course be obtained at higher temperature with a benefic effect on throughput, but at the expense of increased precursor consumption.

The duration of the precursor exposure steps was set by measuring the deposition rate at varying exposure time (Fig. 3.5) and choosing the shortest time that would guarantee saturated reactions. Longer exposures would increase precursor consumption and process time, while too short durations would compromise film uniformity due to incomplete reactions on part of the wafer surface. The duration of the purge steps was determined as the minimum time that would achieve complete purging of the precursors, as observed by both residual gas analysis (RGA) and deposition uniformity. From a practical point of view, slightly longer purge steps than strictly necessary were employed. While this has an impact on throughput, it acts as a guarantee that the precursors are effectively and reliably purged during each run, since incomplete purging of the precursors from the reactor environment would cause a CVD contribution to the film growth, degrading uniformity and reproducibility.

Fully self-saturating reactions were only observed in the temperature range 200–300 °C, where a proper ALD window with a temperature-independent deposition rate of $(1.50 \pm 0.05) \text{Å/cycle}$ was identified (Fig. 3.6), achieving 1-5% uniformity on 200 mm substrates. The higher deposition rate, compared to the value of 0.22 Å/cycle reported in [108], can be explained by the higher reactivity of the Cp ligand compared to thd. The high reactivity is also the likely cause for the very low incubation time of 2-6 cycles observed for deposition on HF-last Si (Fig. 3.7), which
3.3. Characterization of Tm$_2$O$_3$ thin films for CMOS applications

A thorough characterization of the film properties is presented in Paper III. An extensive analysis of the deposited films is important toward actual use of the
developed process, in the envisioned application of TmSiO formation as well as in completely different areas where the availability of a novel process might enable new applications of Tm$_2$O$_3$. In this section, the film properties that are most important for integration in a CMOS process flow are discussed.

**Main deposition parameters** Deposition temperature, cycle length and precursor choice should be reasonable from the point of view of integration and throughput. The developed thermal ALD process works at a relatively low temperature of 200–250°C, which makes it fully compatible with any material that could be present on the wafer surface at the gate deposition step in any CMOS integration scheme. The duration of each deposition cycle can be reduced by increasing the source temperature, at the expense of increased precursor consumption. The precursor itself, at the time of this writing, poses issues of limited availability and relatively high price, but both obstacles would likely disappear if market demand were to grow.

**Film quality** The process was verified to proceed by pure ALD-type growth, which is crucial for obtaining the required uniformity and conformality. The concentration of impurities in the films was measured by time of flight elastic recoil detection analysis (ToF-ERDA), finding low C and H content (< 1 at%) in films deposited at 200°C, while a high C content (~10 at%) was measured in films deposited at 300°C. The C concentration was found to decrease after annealing in N$_2$ at 600°C (Fig. 3.8), which can be explained by the presence of a carbonate phase in the films deposited at higher temperatures and its decomposition during annealing [121]. The films were slightly off-stoichiometric (Tm:O ratio of 1.63:1, compatible with oxygen-rich Tm$_2$O$_3$), but perfect stoichiometry is not critical in this case since the oxide is only used to form a silicate and is later etched away (although the excess oxygen in the film might have an influence on the silicate growth).

**Direct contact** Deposition in direct contact to the Si surface is a desired property for high-k dielectrics for CMOS applications, and is especially needed for processes aiming to replace the interfacial layer. The choice of a highly reactive organometallic precursor allowed the deposition to nucleate on HF-last Si with low incubation time and allowed the use of H$_2$O as oxidizing agent, which avoids the growth of a thick SiO$_x$ layer between Si and Tm$_2$O$_3$ (Fig. 3.9). Care was also taken in loading the samples into the ALD reactor immediately after immersion in diluted HF chemistry, rinsing and drying in order to minimize native oxide regrowth [122] and adsorption of contaminations.
3.3. Characterization of Tm$_2$O$_3$ thin films for CMOS applications

Figure 3.8. Carbon concentration in Tm$_2$O$_3$ films versus deposition temperature, with and without annealing in N$_2$ for 5 min at 600 °C. Increasing C content is observed in films deposited at higher temperature, but acceptable levels can be achieved by low-temperature deposition and/or by annealing.

Figure 3.9. TEM cross-section of 12 nm thick Tm$_2$O$_3$ layer deposited at 300 °C. A graded transition is observed at the interface with Si, while the lack of a clearly visible SiO$_x$ layer indicates that the H$_2$O-based process does not induce strong oxidation of the Si substrate even at the highest temperature in the ALD window. (Paper III)
3.4 Possible process improvements via precursor optimization

Two main issues were identified in the developed process, namely the high C content at 300 °C deposition temperature and the oxygen-rich film stoichiometry. Neither issue was critical in the present application, but stoichiometric deposition could be desirable in other applications, e.g. if Tm$_2$O$_3$ was to be used as bulk high-k oxide, as could the possibility to deposit the oxide at higher temperatures, e.g. to use Tm as doping element in a host oxide (such as HfO$_2$) which is deposited at higher temperatures.

The high C content was strongly linked to temperature-activated precursor decomposition, as supported by RGA measurements, previous reports of partial decomposition in Cp-based lanthanide oxide depositions [123] and theoretical models [124]. Therefore, using an organometallic precursor with higher thermal stability could provide a solution to this problem, enabling deposition of Tm$_2$O$_3$ at higher temperatures with low impurity content, although it is hard to predict if the same approach would provide a stoichiometric deposition. More stable Tm precursors are not commercially available at the time of this writing, but from basic considerations on the coordination chemistry of lanthanide elements it is expected that the use of larger Cp-based ligands should provide volatile compounds with higher thermal stability [114].
Chapter 4

Integration of TmSiO in a CMOS process

The fabrication and characterization of high-k/metal gate MOSFETs has been at the core of this work, since integration in a realistic CMOS process is the only way to verify the suitability of a novel gate stack. Fabrication of long-channel devices is especially relevant to research on gate stacks since it enables meaningful extraction of the relevant device properties, especially channel mobility, without influence from short-channel effects. MOS capacitors have also been employed in much of the work, since the simplified process allows faster process optimization cycles and the devices represent valid test structures for many important properties of the gate stack, such as EOT, flatband voltage and interface state density.

The first section of the present chapter describes the MOSFET fabrication process and the electrical characterization procedure. In the second part, the implementation of TmSiO in a CMOS process and its compatibility with complementary device operation are discussed by analyzing and comparing three different dielectric stacks with similar EOT $\sim 1.2$ nm, i.e. SiO$_x$/HfO$_2$, TmSiO/Tm$_2$O$_3$ and TmSiO/HfO$_2$. The final section describes the results of an experiment attempting to integrate the TmSiO interfacial layer with a higher-k oxide. Integration of TmSiO in Hf-based MOSFETs achieving sub-nm EOT is discussed in the next chapter.

4.1 MOSFET fabrication and characterization

The process employed in this work for MOSFET fabrication is a gate-last high-k/metal gate CMOS process. Compared to the actual industrial integration scheme discussed in Chapter 1, the fabrication followed a simplified process flow due to the lack of in-house CMP equipment, which is necessary for a short-turnaround process. However, the aim of the process was the fabrication of long-channel devices, since they represent optimal test structures for evaluation of novel gate stacks, as long as the design completely suppresses short-channel effects. Therefore, a simplified
4. Integration of TmSiO in a CMOS process

gate-last process flow which is restricted to long-channel devices was implemented in this work.

In the following, the employed process is described and deviations from standard process flows are discussed.

4.1.1 Well implantation and isolation

The substrates used in this work consist in bulk (100) Si 100 mm wafers, doped with B at a resistivity of 20–40 Ω cm (corresponding to a doping density of ~4 × 10^{14} cm^{-3}). While low-resistivity substrates with lowly doped epitaxial layers are more common in advanced processes, the lack of circuits and therefore the lack of concerns on latchup made it reasonable to use more easily available wafers. Intrinsic gettering was employed to reduce the concentration of impurities in the active device region.

A twin-well process was used, as common in CMOS technology, where nwells and pwells were doped at symmetric levels of 10^{17} cm^{-3}. The doping density is sufficiently high to completely suppress short-channel effects in L_G=3 µm devices. A long drive-in furnace annealing (9 hours in N_2 at 1150°C) was performed to ensure complete dopant activation, full defect annealing and deep wells.

Isolation between adjacent devices is usually accomplished by shallow trench isolation (STI) in industrial processes due to its advantages in terms of increased packing density and planar topography. Given the wide separation among devices in the maskset design and the use of I-line lithography with sufficient depth of focus, the simpler localized oxidation of silicon (LOCOS) isolation technique was used in this work.

4.1.2 Source/drain implantation

Conventional source/drain implantation is a viable option for long-channel devices, since relatively deep source/drain regions can be used, thereby minimizing the impact of source/drain resistance on drain current without requiring the integration of raised source/drain technology.

Source and drain regions were implanted using a resist mask, which takes on the role of dummy gates in RMG processes. The implantation was performed through a screen oxide to minimize channeling. High implantation dose and energy were used, so that a doping density of ~10^{21} cm^{-3} and a ~80 nm projected range were obtained. This ensured negligible source/drain resistance for long-channel device operation and obviated the need for extended silicidation of the source/drain areas. After removal of the resist mask, dopant activation was performed by RTA in N_2 for 10 s at 1050°C.

4.1.3 Gate stack

As mentioned before, a gate-last integration scheme was employed, but it is worth mentioning that a self-aligned implementation, such as the RMG process described
4.1. MOSFET fabrication and characterization

in Chapter 1, was not followed. This makes the process incompatible with short-channel devices but perfectly viable and easier to manufacture for the long-channel devices used in this work.

After source/drain activation and removal of sacrificial oxide, the gate dielectric stack was deposited. Since different dielectric stacks were used during this work, the reader is referred to the description of the different experiments in the appended papers for details on the deposition process. Generally speaking, the following dielectrics were employed in this work:

- $\text{Al}_2\text{O}_3$, deposited by ALD from trimethylaluminum and water at 200–350°C;
- $\text{HfO}_2$, deposited by ALD from bis(methylcyclopentadienyl)methoxymethyl-hafnium and water at 350°C;
- $\text{La}_2\text{O}_3$, deposited by ALD from tris(isopropylcyclopentadienyl)lanthanum at 250–350°C;
- $\text{LaSiO}$, formed by reaction of $\text{La}_2\text{O}_3$ with Si (see Chapter 2).
- $\text{SiO}_2$, deposited by PECVD from silane and nitrous oxide;
- $\text{SiO}_x$, grown by chemical oxidation in the ALD reactor (see Section 4.2);
- $\text{Tm}_2\text{O}_3$, deposited by ALD from tris(cyclopentadienyl)thulium and water at 200–250°C;
- $\text{TmSiO}$, formed by reaction of $\text{Tm}_2\text{O}_3$ with Si (see Chapter 2).

TiN was always used as workfunction metal in the MOSFET process and was deposited by reactive sputtering or ALD (at 350°C using $\text{TiCl}_4$ and $\text{NH}_3$ as precursors) to a thickness of 15 nm. The value was chosen because a thickness of 10–20 nm has been reported to minimize the effect of TiN-induced strain on channel mobility, making a comparison to reported mobility values more meaningful [125]. A thicker (100 nm) TiW layer was then deposited by sputtering to enable well-controlled and vertical gate etch. The gate was patterned using a mask with intentional overlap to the source/drain implantation mask in order to compensate overlay errors in the stepper, although care was also taken in minimizing misalignments by using automatic and manual corrections in the exposure pass.

4.1.4 Passivation and metallization

Passivation of the devices was performed by deposition of PECVD dielectrics, which were then patterned to provide contact holes to source, drain, gate and well regions. A $\text{SiO}_2$/SiN stack (30/70 nm) was initially employed, but a newer passivation and contact etch process was later integrated, providing 400 nm SiO$_2$ passivation with the advantage of reduced parasitic pad capacitance.
A single level of metallization, consisting in a PVD Ti/TiW/Al stack, was employed, since only individual devices were fabricated. A 10 nm Ti layer was used to ensure low contact resistivity toward the TiN/TiW gates, the highly doped source/drain areas and the highly doped well contact areas (which were implanted at the same time as the nFET and pFET source/drain implantations). A 100 nm TiW layer served to provide good step coverage in the contact holes and to act as Al diffusion barrier. A 100–500 nm Al layer was used to ensure good electrical contact with metallic probes for electrical characterization. 100 × 100 µm² pads were patterned in the metal stack to be able to contact the devices on a wafer probestation.

As final step, FGA was performed in 10% H₂ in N₂ at 400°C for 30 min, unless otherwise stated.

4.1.5 MOS capacitors

A good part of process optimization and evaluation has been performed on MOS capacitors. While such structures can not be used to analyze carrier transport in the channel, they can still provide useful quantitative information on the gate stack, especially regarding EOT, flatband voltage and interface state density. MOS capacitors can also be used to compare the effects of different treatments on leakage current, but it should be kept in mind that the current density values can not be treated as an estimate of the gate leakage current density in a MOSFET employing the same gate stack, as current in a MOS capacitor with thin gate oxide is strongly limited by transport in the bulk semiconductor [100].

The use of MOS capacitors can effectively speed up the process optimization phase, since they require a much simpler process compared to MOSFETs. MOS capacitors also provide an easy way to test unconventional materials which would be difficult to integrate in a full CMOS process. For example, lift-off patterning of Ag, Cr and Pt has been used to fabricate capacitors aimed at testing the compatibility of the TmSiO/HfO₂ dielectric stack with gate-last workfunction tuning techniques (see Section 4.3).

The process used for the fabrication of MOS capacitors varied widely depending on the requirements of the specific experiments, therefore the reader is referred to the appended papers for detailed process descriptions.

4.1.6 Electrical characterization and analysis

Electrical characterization of the fabricated devices has provided most of the results presented in this thesis.

Electrical measurements were performed using an automated characterization setup for wafer-level probing and a manual setup for single measurements on wafers or small samples. The automated setup consists in a Cascade 12000 semi-automatic wafer prober, connected to a Keithley SCS 4200 for DC, AC and pulsed measurements. The setup allows automatic mapping of wafers and uploads the
measured data to a SQL database, which was designed and implemented by the
author together with MATLAB classes for data access. The manual setup consists in
a Cascade 11000 wafer prober connected to a Keithley SCS 4200 for current-voltage
(IV) measurements and to a HP 4284A for capacitance-voltage (CV) measurements.
Analysis of the measured data was mostly performed by standard techniques,
which are described in reference texts such as [126] and [127]:

- threshold voltage, subthreshold slope and leakage current density were ex-
  tracted from DC IV measurements;
- channel mobility was measured by the split-CV technique;
- the quality of the interface was assessed by the conductance technique on
  MOS capacitors and by charge pumping on MOSFETs;
- BTI was measured under constant voltage stress (CVS) conditions;
- TDDB was measured under voltage ramp stress (VRS) conditions [128].

The analysis procedure for CV measurements deserves separate treatment, given
its importance in the assessment of novel gate stacks. CV measurements on MOS
 capacitors and gate-to-bulk CV curves on MOS transistors were analyzed with the
models proposed by Ghibaudo [129] and Hauser [130] to extract flatband voltage
and EOT. The extracted values on series of capacitors with varying thickness were
also used to extract the effective workfunction according to [131]. To ensure correct
application of the CV models, series resistance effects were preliminarily corrected
by fitting impedance values measured at multiple frequencies to a series-parallel RC
model, as an extension of the two-frequency technique described in [132,133]. In the
case of MOS capacitors measured on small samples, the effect of contact parasitic
capacitance was quantified and corrected by measuring structures with varying area.

4.2 MOSFET characteristics with TmSiO interfacial layer

The basic characterization of thulium silicate formation and its compatibility with
CMOS integration schemes has been presented in Chapter 2, where TmSiO was
shown to outperform the more commonly employed LaSiO in the envisioned in-
tegration process. This section discusses the integration of TmSiO as interfacial
layer in high-k/metal gate MOSFETs and compares the resulting performance to
the conventional chemical oxide interfacial layer. To this aim, a baseline SiO$_x$ IL
process was first developed and characterized and then a comparison with the novel
TmSiO IL was performed.

4.2.1 SiO$_x$ interfacial layer

The most common way to grow a chemical oxide interfacial layer for high-k/metal
gate stacks is to oxidize the Si substrate in a DI-H$_2$O bath with low O$_3$ content
4. Integration of TmSiO in a CMOS process

Figure 4.1. EOT versus IL thickness for three different in situ oxidations (O$_3$/O$_2$, O$_3$/O$_2$/H$_2$O and pulsed), compared to low-temperature oxidation in SC1 and etched back thermal oxide. Sub-8Å physical thickness can be obtained by the O$_3$ and pulsed methods. (Paper I)

at room or low temperature [134]. Given the lack of a specialized wet bench, a novel way to conduct chemical oxidation of Si has been developed in this work, consisting in the exposure to a diluted O$_3$ gas flow in the ALD reactor. The reactor temperature was set equal to the deposition temperature of HfO$_2$ (350°C) in order to be able to grow the interfacial layer in situ, minimizing the exposure of the interface to contaminations. Compared to the common room-temperature chemical oxidation, the higher temperature employed by this method could in principle be beneficial to the quality of the interface with Si, while still being low enough that a thickness target of ~0.5 nm could be achieved by using a low O$_3$ flow diluted in O$_2$, H$_2$O and Ar.

Several variants of the procedure were tested and compared to conventional room-temperature oxidation in SC-1 (H$_2$O:NH$_4$OH:H$_2$O$_2$ 5:1:1) and to etched back thermal oxide, finding that sub-nm interfacial layers could be grown by continuous exposure to an O$_3$/O$_2$ flow or by alternating a O$_3$/O$_2$/H$_2$O flow with an inert Ar flow (Fig. 4.1). The development and characterization of the method are discussed in Paper I, where the SiO$_x$ interfacial layer was also integrated in HfO$_2$/TiN MOSFETs, showing well-behaved IV characteristics and acceptable channel mobility.

4.2.2 TmSiO interfacial layer

The primary comparison between the novel TmSiO interfacial layer and the conventional chemical oxide consists in the scaling capabilities. Chemical oxides can be scaled below 0.3–0.4 nm by means of scavenging, but at the cost of strongly degraded device properties (see Section 1.3). On the other hand, TmSiO was shown in Chapter 2 and Paper II to contribute values of EOT below 0.3 nm while keeping similar quality to thicker films. However, the IL EOT was in that case estimated
from TmSiO/SiO$_2$ stacks, while it is crucial that the same result can be achieved when integrated with a bulk high-k dielectric.

To this aim, the TmSiO interfacial layer was integrated in MOS capacitors with three different bulk oxides, namely SiO$_2$, Tm$_2$O$_3$ and HfO$_2$. In the first stack, the excess Tm$_2$O$_3$ was etched and PECVD SiO$_2$ was deposited as bulk oxide, with the purpose of using a stable and well-characterized material to facilitate the extraction of the intrinsic properties of TmSiO. In the second stack, Tm$_2$O$_3$ was kept as bulk oxide, with the purpose of verifying whether the quality of the TmSiO layer was affected by the etching procedure. In the third stack, the excess Tm$_2$O$_3$ was removed and HfO$_2$ was deposited as bulk oxide, with the purpose of verifying the compatibility with a realistic, industrially-relevant material. TiN was used as gate metal in all cases. The IL EOT was extracted by fabricating several samples with varying thickness of the bulk oxides (Fig. 4.2) and fitting the data to the equation:

$$EOT = EOT_{IL} + EOT_{ox} = EOT_{IL} + \frac{k_{SiO_2}}{k_{ox}} t_{ox}$$

where $EOT_{IL}$ is the contribution of the interfacial layer to the total EOT, $EOT_{ox}$ is the contribution of the bulk oxide to the total EOT, $k_{SiO_2}$ and $k_{ox}$ are the relative dielectric constants of SiO$_2$ and the bulk oxide respectively, and $t_{ox}$ is the physical thickness of the bulk oxide. For all three data sets, the EOT of the IL lies in the range (0.25 ± 0.15) nm, leading to the important conclusion that the properties of the interfacial layer are not affected by the etching-deposition procedure, which is crucial to its integrability. The extracted values are also strongly competitive toward conventional chemical oxide interfacial layers and justify the integration of TmSiO in sub-nm HfO$_2$/TiN MOSFETs, which is discussed in Chapter 5.

The TmSiO interfacial layer was then compared to the in situ SiO$_x$ process in terms of quality of the interface with Si. Fig. 4.3a reports the interface state
density measured by the conductance method. The slightly lower values measured on the gate stack employing SiO$_x$ correlate with the higher extracted IL EOT (0.5 nm) compared to TmSiO (0.2 nm). As a matter of fact, the measured values of interface state density for gate stacks employing TmSiO interfacial layers with different estimated IL EOT follow the same trend as values reported for scaled chemical oxide IL [135] (Fig. 4.3b).

Finally, TmSiO was integrated in MOSFETs using Tm$_2$O$_3$ as bulk oxide and TiN as gate metal, yielding well-behaved IV characteristics (Fig. 4.4) with low subthreshold slopes of $\sim$70 mV/dec, which is very close to what can be calculated from a well doping density of $10^{17}$ cm$^{-3}$ and a CET of $\sim$1.6 nm. Similar CET values had been obtained in the SiO$_x$/HfO$_2$ process, enabling direct comparison of channel mobility, with the result that the TmSiO/Tm$_2$O$_3$ stack provides 20% higher mobility than the SiO$_x$/HfO$_2$ stack (Fig. 4.5). Such result can be explained by the higher physical thickness of the high-k IL, as will be shown in Chapter 5 from the analysis of devices employing a TmSiO/HfO$_2$/TiN gate stack at sub-nm EOT.

4.3 EWF control in TmSiO/HfO$_2$ gate stacks

Threshold voltage ($V_T$) control is a crucial issue in high-k/metal gate CMOS technology. The traditional way to achieve symmetric threshold voltages, via dual-doped poly-Si and channel implantation, can not be applied, since poly-Si gates are incompatible with high-k dielectrics and high channel doping is incompatible
4.3. EWF control in TmSiO/HfO$_2$ gate stacks

**Figure 4.4.** Normalized drain current versus gate voltage for TmSiO/Tm$_2$O$_3$ MOSFETs with $L_G=3 \, \mu m$ and $W_G=50 \, \mu m$. 100 nFETs and 100 pFETs have been measured over a 100 mm wafer. (Paper IV)

**Figure 4.5.** Channel mobility in TmSiO/Tm$_2$O$_3$ MOSFETs. 20% higher values than SiO$_x$/HfO$_2$ devices are measured, in line or slightly higher than literature trends. (Paper IV)
4. Integration of TmSiO in a CMOS process

Figure 4.6. $I_D$-$V_G$ characteristics from SiO$_x$/HfO$_2$, TmSiO/Tm$_2$O$_3$ and TmSiO/HfO$_2$ MOSFETs. 30-100 devices have been measured for each polarity and for each gate stack. The higher ($V_{TN}+|V_{TP}|$) in SiO$_x$/HfO$_2$ devices is due to the higher well doping density which was implanted in that batch.

with novel undoped devices (e.g. FinFETs and SOI) and with the need to minimize scattering and variability sources. The most commonly used techniques for threshold voltage control have been described in Chapter 1 and consist in the use of dual metal gates in gate-last integration schemes and in the use of dual capping layers with a single metal electrode in gate-first schemes.

Therefore, compatibility of the novel TmSiO interfacial layer with at least one threshold voltage control technique is necessary toward its possible integration in a future technology node. Such investigation has been the subject of Paper V, where integration of TmSiO in realistic HfO$_2$-based gate stacks has been shown to preserve compatibility with both threshold voltage control techniques. Preliminary evidence of the potential compatibility of TmSiO with complementary device operation is shown in Fig. 4.6, which compares $I_D$-$V_G$ characteristics from MOSFETs using three different gate stacks: SiO$_x$/HfO$_2$/TiN, TmSiO/Tm$_2$O$_3$/TiN and TmSiO/HfO$_2$/TiN. A quantitative comparison is hindered by the fact that the batch with the SiO$_x$/HfO$_2$ stack had a higher well doping and that the TmSiO/HfO$_2$ stack achieved lower EOT (0.6–1 nm versus 1–1.4 nm for the two other stacks), however a qualitative comparison can be performed. The TmSiO/Tm$_2$O$_3$ stack shows a pronounced shift toward negative voltages, which is expected from the use of Tm$_2$O$_3$ as bulk oxide, as many lanthanide oxides have been reported to be only compatible with nFET operation [13]. On the other hand, the TmSiO/HfO$_2$ and SiO$_x$/HfO$_2$ stacks present similar threshold voltages, indicating that the substitution of chemical oxide with thulium silicate has little influence on EWF.

Such conclusion is confirmed by the extracted EWF on SiO$_x$/HfO$_2$ and TmSiO/HfO$_2$ MOS capacitors, which show similar EWF tunability (Fig. 4.7). As a consequence, the flatband voltage of the stacks can be effectively set by a
4.4 Integration of TmSiO IL with a higher-k bulk oxide

proper choice of the gate metal (Fig. 4.8), therefore symmetric threshold voltages
should be easily achievable in gate-last integration schemes, which are becoming
dominant in the industry thanks to their better resilience to flatband voltage roll-off.

Compatibility with gate-first $V_T$ control techniques was also addressed, showing
that the flatband voltage can be shifted by integration of La$_2$O$_3$ or Al$_2$O$_3$ capping
layers, although shifts larger than $\sim$150 mV can only be obtained at the cost of
undesirable EOT increases. This is in contrast to the conventional SiO$_x$/HfO$_2$
system, where the Al$_2$O$_3$ capping layer gives an undesirable EOT increase, but the
La$_2$O$_3$ capping layer gives a desirable EOT decrease [43]. The difference lies in
the fact that, in the conventional SiO$_x$/HfO$_2$ stack, diffusion of La atoms to the
interfacial layer results in a partial silicate formation, so that there is an optimum
point where the additional EOT from the La$_2$O$_3$ layer is more than compensated
by the increased dielectric constant of the interfacial layer. In the TmSiO process,
instead, the interfacial layer is already a silicate, therefore no further increase in
dielectric constant is expected.

Comparing the two technologies for silicate formation (i.e. direct TmSiO for-
mation and La diffusion from La$_2$O$_3$ caps), the integration of TmSiO as interfacial
layer provides the following advantages:

- it is formed before the rest of the gate stack, without requiring additional
capping layers or high temperature anneals, thereby being compatible with
both gate-last and gate-first integration schemes;

- its composition and dielectric constant are set by the formation temperature,
while La diffusion to the SiO$_x$ interfacial layer is believed to only increase its
dielectric constant slightly;

- it is compatible with both nFET and pFET operation, while the La$_2$O$_3$ capping
layer induces a flatband voltage shift which is only favorable for nFETs.

4.4 Integration of TmSiO IL with a higher-k bulk oxide

Sections 4.2 and 4.3 showed integration of TmSiO as interfacial layer in high-
k/metal gate MOSFETs employing Tm$_2$O$_3$ as bulk oxide and in MOS capacitors
employing HfO$_2$ as bulk oxide. The encouraging results motivated the fabrication
and optimization of TmSiO/HfO$_2$ MOSFETs (addressed in Chapter 5), achieving
sub-nm EOT with enhanced channel mobility. While this stack proved competitive
to the conventional SiO$_x$/HfO$_2$ stack and could provide better short-term scalability,
a longer-term solution would be the integration of the high-k TmSiO interfacial
layer with a higher-k bulk oxide. Such cointegration has been tested in a separate
experiment and is described here.

Crystalline HfO$_2$ has been used as higher-k oxide in this experiment, obtained
through doping of the host HfO$_2$ oxide with Al and annealing in N$_2$ [77]. The
process has first been reproduced using SiO$_x$ as interfacial layer. Since a regrowth
4. Integration of TmSiO in a CMOS process

Figure 4.7. Effective workfunction in TmSiO/HfO$_2$ and SiO$_x$/HfO$_2$ gate stacks employing different workfunction metals. Integration of TmSiO does not affect EWF tunability. (Paper V)

Figure 4.8. Flatband voltage versus EOT for TmSiO/HfO$_2$ gate stacks employing different workfunction metals. The flatband voltage can be effectively set by the gate metal in a gate-last process. (Paper V)
4.4. Integration of TmSiO IL with a higher-k bulk oxide

Figure 4.9. Flatband voltage versus EOT for TmSiO/HfO$_2$ gate stacks employing different capping layers. The flatband voltage can be adjusted by integration of a dielectric capping layer in a gate-first process, although a trade-off with EOT is observed. (Paper V)

of the IL can be assumed from the results in [77], the main process parameters have been varied in order to obtain an increased dielectric constant while minimizing the IL regrowth. Additionally, the cycle ratio between HfO$_2$ and Al$_2$O$_3$ has been adapted to take into account the different deposition rate achieved by the employed bis(methylcyclopentadienyl)methoxymethylhafnium precursor. The resulting factorial experiment is shown in Table 4.1, where for each experimental condition several samples have been fabricated with varying thickness of the bulk oxide in order to extract IL EOT and k$_{HfO_2}$ independently. The results of the experiment indicate that a dielectric constant $\sim$30, compatible with cubic or tetragonal HfO$_2$, can be obtained with a doping ratio of 28:1 and annealing at 800°C. Short (1 s) anneals are effective in increasing the dielectric constant and are beneficial in terms of a lower IL regrowth, compared to longer (30 s) treatments.

In the second part of the experiment, the process was repeated on TmSiO interfacial layers, fixing the cycle ratio at 28:1 and the annealing time at 1 s, according to the results of the first experiment. Annealing temperature was kept as a variable, given its effect on IL EOT and given the possibility that Tm diffusion from the IL might further lower the crystallization temperature. Additionally, two types of annealings were considered, pre- and post-metal, since annealing through a thick TiN layer would prevent oxygen content in the N$_2$ atmosphere from reaching the interfacial layer, although this modification would reduce the compatibility with standard CMOS integration schemes. The results (Table 4.2) indicate that annealing at 800 °C is necessary to obtain a dielectric constant which is significantly higher than monoclinic or amorphous HfO$_2$, while annealing after the deposition of the gate metal can reduce the IL regrowth.

However, regrowth of the interfacial layer was always observed, so that the
4. Integration of TmSiO in a CMOS process

<table>
<thead>
<tr>
<th>Cycle ratio</th>
<th>Annealing temperature</th>
<th>Annealing time</th>
<th>IL EOT</th>
<th>k_{HfO_2}</th>
</tr>
</thead>
<tbody>
<tr>
<td>28:1</td>
<td>700 °C</td>
<td>1 s</td>
<td>0.81 nm</td>
<td>25.5</td>
</tr>
<tr>
<td>28:1</td>
<td>700 °C</td>
<td>30 s</td>
<td>1.00 nm</td>
<td>25.8</td>
</tr>
<tr>
<td>28:1</td>
<td>800 °C</td>
<td>1 s</td>
<td>0.95 nm</td>
<td>30.4</td>
</tr>
<tr>
<td>28:1</td>
<td>800 °C</td>
<td>30 s</td>
<td>1.30 nm</td>
<td>31.0</td>
</tr>
<tr>
<td>34:1</td>
<td>700 °C</td>
<td>1 s</td>
<td>0.82 nm</td>
<td>23.4</td>
</tr>
<tr>
<td>34:1</td>
<td>700 °C</td>
<td>30 s</td>
<td>1.03 nm</td>
<td>23.0</td>
</tr>
<tr>
<td>34:1</td>
<td>800 °C</td>
<td>1 s</td>
<td>0.90 nm</td>
<td>28.3</td>
</tr>
<tr>
<td>34:1</td>
<td>800 °C</td>
<td>30 s</td>
<td>1.25 nm</td>
<td>28.0</td>
</tr>
</tbody>
</table>

Table 4.1. Analysis of the factorial experiment aiming to integrate the higher-k c-HfO_2 bulk oxide on top of the SiO_x IL.

<table>
<thead>
<tr>
<th>Annealing temperature</th>
<th>Annealing type</th>
<th>IL EOT</th>
<th>k_{HfO_2}</th>
</tr>
</thead>
<tbody>
<tr>
<td>700 °C</td>
<td>pre-metal</td>
<td>0.79 nm</td>
<td>24.6</td>
</tr>
<tr>
<td>700 °C</td>
<td>post-metal</td>
<td>0.55 nm</td>
<td>23.2</td>
</tr>
<tr>
<td>800 °C</td>
<td>pre-metal</td>
<td>0.92 nm</td>
<td>29.7</td>
</tr>
<tr>
<td>800 °C</td>
<td>post-metal</td>
<td>0.60 nm</td>
<td>29.3</td>
</tr>
</tbody>
</table>

Table 4.2. Analysis of the factorial experiment aiming to integrate the higher-k c-HfO_2 bulk oxide on top of the TmSiO IL.

EOT decrease caused by the higher dielectric constant is offset by the increased IL EOT (Fig. 4.10), leading to no net advantage of this process over the simpler TmSiO/HfO_2 process. Successful co-integration of TmSiO and crystalline HfO_2 may still be possible if IL regrowth could be suppressed via an optimized annealing process, e.g. by flash annealing.
4.4. Integration of TmSiO IL with a higher-k bulk oxide

Figure 4.10. EOT versus bulk oxide thickness for TmSiO/Hf$_x$Al$_y$O/TiN gate stacks with and without annealing (performed after the metal gate deposition in N$_2$ at 800 °C for 1 s). The increased dielectric constant is compensated by a 2x increase in IL EOT, leading to the conclusion that the process has worse scalability than the simpler SiO$_x$/HfO$_2$ process.
Chapter 5

Enhanced channel mobility at sub-nm EOT

In the previous chapters, thulium silicate has been demonstrated as a plausible candidate for integration as high-k interfacial layer in high-k/metal gate CMOS technology. In Chapter 2, the silicate formation was analyzed and a suitable integration process flow, compatible with both gate-first and gate-last integration schemes, was presented. In Chapter 4, the TmSiO interfacial layer was implemented in a gate-last CMOS process, providing competitive MOSFET characteristics in terms of IL EOT, interface state density, threshold voltage control and channel mobility.

This chapter demonstrates integration of TmSiO in an industrially relevant gate stack, employing HfO$_2$ as bulk high-k oxide and achieving EOT as low as 0.6 nm. It is shown that TmSiO may indeed be a valuable material for integration in future (sub-10 nm) technology nodes, providing an improved EOT-mobility trade-off, with ~20% higher channel mobility compared to state-of-the-art SiO$_x$/HfO$_2$ gate stacks. Optimization of the annealing conditions for the TmSiO/HfO$_2$ stack is then addressed and the effect of annealing on gate leakage current, interface quality and threshold voltage is discussed. Afterwards, the suitability of the stack from the point of view of reliable device operation is addressed, followed by an analysis of the effect of the high-k IL on scattering mechanisms and channel mobility.

5.1 Sub-nm-EOT TmSiO/HfO$_2$ MOSFETs

The encouraging results outlined in Chapter 4 showed that TmSiO can be integrated as interfacial layer in a generic gate stack and provide competitive advantages over the common chemical oxide IL. Therefore, the next step was to integrate TmSiO in a realistic Hf-based gate-last CMOS process, in order to evaluate its compatibility with the high-k dielectric which is today most relevant in the manufacturing environment.

The process followed the general description given in Chapter 4 and integrated a
5. Enhanced channel mobility at sub-nm EOT

Figure 5.1. (a) Process flow and (b) TEM cross-section of the TmSiO/HfO$_2$/TiN gate stack implemented in gate-last MOSFETs. (Paper VI)

bilayer gate dielectric consisting of TmSiO as interfacial layer and HfO$_2$ as bulk high-k oxide (Fig. 5.1). In the first demonstration of sub-nm-EOT operation, presented in Paper VI, no post deposition anneal (PDA) was performed before deposition of the TiN gate electrode, and a standard (30 min at 400 °C) FGA was employed (see Section 5.2 for a discussion on the effects of annealing).

The TmSiO/HfO$_2$ dielectric stack achieved EOT as low as 0.65 nm on nFETs (Fig. 5.2a) and 0.8 nm on pFETs (Fig. 5.2b). The values were obtained by fitting the measured gate-to-bulk CV curves using the widely employed CVC fitting tool [130]. The software is not a full quantum-mechanical simulator, which would in principle be required for proper determination of EOT in scaled gate stacks, but achieves reasonably good agreement with quantum-mechanical simulators by considering first-order corrections to the classical MOS theory [136]. The use of CVC in this work is motivated by the need to compare the results to literature values and is reinforced by the excellent agreement with the extracted CET, which was not obtained by fitting but was directly calculated from the measured gate-to-channel capacitance at $|V_G|=1.5$ V (Table 5.1). The ~0.4 nm difference between EOT and CET is commonly observed in Si MOSFETs with metal gates and is explained as the contribution of the vertical distribution of carriers in the channel, which does not peak at the interface [137]. The obtained EOT values are also consistent with the gate stack design, since TmSiO formed at 500–550 °C provides an IL EOT of ~0.25 nm (see Chapter 4) and a (2.0 ± 0.5) nm thick HfO$_2$ layer was deposited, yielding an EOT contribution of (0.45 ± 0.10) nm (assuming a dielectric constant of 18, which is the value that was extracted during the initial process development phase of HfO$_2$ deposition by ALD).

$I_D$-$V_G$ characteristics for nFETs and pFETs are shown in Fig. 5.3. Excel-
5.1. Sub-nm-EOT TmSiO/HfO$_2$ MOSFETs

![Figure 5.2](image)

**Figure 5.2.** Gate-to-bulk ($C_{GB}$) and gate-to-channel ($C_{GC}$) capacitance-voltage characteristics measured on TmSiO/HfO$_2$ (a) nFETs and (b) pFETs. 30 devices for each polarity are shown, together with CVC fitting of the $C_{GB}$ curves. Negligible hysteresis was observed in the CV curves (not measurable in pFETs and in most nFETs, with a maximum of 50 mV in some nFETs). (Paper VI)

<table>
<thead>
<tr>
<th>Type of device</th>
<th>EOT (nm)</th>
<th>CET (nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>nFET</td>
<td>0.65-1.1</td>
<td>1.0-1.6</td>
</tr>
<tr>
<td>pFET</td>
<td>0.8-1.2</td>
<td>1.25-1.6</td>
</tr>
</tbody>
</table>

**Table 5.1.** Values of EOT and CET extracted from CV characteristics of TmSiO/HfO$_2$ MOSFETs.

Lent uniformity is observed over the wafer, and the low values of subthreshold slope (65–70 mV/dec for pFETs) are consistent with the interface state density $< 2 \times 10^{11}$ cm$^{-2}$eV$^{-1}$ measured on MOS capacitors in Chapter 2. The higher subthreshold slopes measured on nFETs are not due to a degraded interface, but rather to high junction leakage. As a matter of fact, TmSiO/HfO$_2$ nFETs achieved similarly low subthreshold slopes of 65–75 mV/dec before FGA (see Fig. 5.17 in Section 5.3), therefore the increased values after FGA are probably due to the non-optimized source/drain dopant activation anneal, leading to defect formation and/or growth during FGA.

The main result from the integration of TmSiO in sub-nm-EOT MOSFETs is shown in Fig. 5.4-5.5, where the high-field electron and hole mobility is compared to literature data on scaled SiO$_x$/HfO$_2$ MOSFETs [43, 44]. TmSiO/HfO$_2$ devices are shown to outperform SiO$_x$/HfO$_2$ devices at the same EOT by 20% for nFETs and 15% for pFETs, leading to the conclusion that integration of TmSiO as interfacial layer can provide a net advantage in the trade-off between EOT and mobility. Further interpretation of the mobility data and a discussion on the influence of the high-k interfacial layer on scattering mechanisms are presented in Section 5.5.
5. Enhanced channel mobility at sub-nm EOT

Figure 5.3. $I_D-V_G$ characteristics measured on TmSiO/HfO$_2$ nFETs and pFETs. 30 devices have been measured for each channel polarity, mapped over a 100 mm wafer. (Paper VI)

Figure 5.4. High-field electron mobility versus EOT, measured on TmSiO/HfO$_2$ nFETs at an effective field of 1 MV/cm, compared to state-of-the-art SiO$_x$/HfO$_2$ devices. (Paper VI)
5.2 Optimized annealing conditions for the TmSiO/HfO$_2$ stack

The TmSiO/HfO$_2$ devices analyzed in the previous section showed excellent properties in terms of EOT and channel mobility but presented the problem of high gate leakage current density ($\sim 10$ A/cm$^2$).

However, further analyses led to the conclusion that the primary cause of the high gate leakage current was not the integration of the high-k IL per se, but rather the lack of proper annealing conditions. Indeed, Fig. 5.6 shows a comparison of MOSFETs with the three different gate stacks analyzed in this work (SiO$_x$/HfO$_2$, TmSiO/Tm$_2$O$_3$ and TmSiO/HfO$_2$), all of them lacking any specific PDA after the bulk high-k deposition (although Tm$_2$O$_3$ underwent the silicate formation anneal). It can be observed that integration of TmSiO has a benefic effect on gate leakage current, since the TmSiO/HfO$_2$ stack has lower gate current density than the SiO$_x$/HfO$_2$ stack even though it achieves lower EOT. Additionally, the TmSiO/Tm$_2$O$_3$ stack provides the lowest gate current density among all the stacks fabricated in this work and shows an exponential trend with respect to EOT, similar to what has been reported for optimized SiO$_x$/HfO$_2$ gate stacks [44, 134]. The comparison indicates that gate current in the TmSiO/Tm$_2$O$_3$ stack is dominated by a tunneling mechanism, which would explain the exponential dependence on EOT, while the higher gate current in HfO$_2$-based gate stacks is explained by the presence of an additional conduction mechanism with a shallower thickness dependence, e.g. a defect-assisted carrier transport mechanism.

Carrier transport via defect bands has been identified as an important conduction
5. Enhanced channel mobility at sub-nm EOT

Figure 5.6. Gate current density versus EOT for SiO$_x$/HfO$_2$, TmSiO/Tm$_2$O$_3$ and TmSiO/HfO$_2$ nFETs and pFETs, measured at 1 V gate bias. Gate current in HfO$_2$-based stacks shows a shallow dependence on EOT, while an exponential dependence is observed in TmSiO/Tm$_2$O$_3$ devices, qualitatively similar to that observed in state-of-the-art SiO$_x$/HfO$_2$ devices.

mechanism in high-k dielectrics and is typically attributed to a high density of oxygen vacancies [138]. The presence of oxygen vacancy defects in HfO$_2$ has been predicted by theoretical models [9] and evidence for extended defect bands has been provided experimentally [55]. Since the concentration of oxygen vacancies can be effectively controlled by means of annealing in a mildly oxidizing environment [139], the development of optimal annealing conditions for the TmSiO/HfO$_2$ stack was considered the most promising approach to reduce the gate leakage current.

Gate current reduction by combined PDA and FGA has been reported in Paper VII. Here, the process development phase is described in detail and some additional results are shown.

5.2.1 PDA development

Different annealing conditions were experimented on TmSiO/HfO$_2$ MOS capacitors and evaluated in terms of EOT, gate current density, interface state density and flatband voltage, with the aim of reducing the density of electrically active defects without causing regrowth of the interfacial layer or degradation of the interface with Si.

Since SiO$_x$/HfO$_2$ stacks are commonly annealed in an atmosphere with low oxygen partial pressure [139], the first treatment to be tested was RTA in N$_2$
5.2. Optimized annealing conditions for the TmSiO/HfO\textsubscript{2} stack

Figure 5.7. Gate current density versus EOT for TmSiO/HfO\textsubscript{2} MOS capacitors with different PDA conditions. (Paper VII)

environment with low O\textsubscript{2} flow. The PDA temperature was set to 550°C, equal to the TmSiO formation temperature, since annealing at higher temperatures was expected to cause strong IL regrowth. As shown in Fig. 5.7, the treatment was found to be ineffective on the TmSiO/HfO\textsubscript{2} stack, since it leads to increased EOT without any beneficial effect on leakage current density, which is actually degraded with respect to the Control wafer (which had the same gate stack deposition process but no PDA).

The second treatment consists in exposure to a dilute O\textsubscript{3} flow in the ALD reactor. A mildly oxidizing environment was obtained by setting the power of the ozone generator to 20% and further diluting the O\textsubscript{3}/O\textsubscript{2} gas mixture in inert Ar flow. The use of O\textsubscript{3} is justified by the decision to set the PDA temperature to 350°C, equal to the HfO\textsubscript{2} deposition temperature, in order to avoid long waiting times after deposition. Therefore, the use of weak oxidizers such as H\textsubscript{2}O or O\textsubscript{2} was expected to produce no effect, since the stack was already exposed to H\textsubscript{2}O at 350°C during the deposition. The Ozone treatment was effective on the TmSiO/HfO\textsubscript{2} dielectric stack, leading to a 10x decrease in gate current density (Fig. 5.7) without any penalty on EOT or interface state density.

A third treatment was also tested, i.e. exposure to direct O\textsubscript{2} plasma at room temperature. However, this treatment led to a strong increase in EOT (~3 nm), therefore it was excluded from the experiment.

5.2.2 Optimization of the ozone PDA

Annealing in O\textsubscript{3}/O\textsubscript{2}/Ar gas flow in the ALD reactor at 350°C provided the best results among the tested treatments, thus the next development phase focused on
5. Enhanced channel mobility at sub-nm EOT

Figure 5.8. (a) Gate current density and (b) interface state density at flatband condition versus EOT for TmSiO/HfO$_2$ MOS capacitors annealed in diluted ozone at different positions in the HfO$_2$ deposition sequence.

optimization of the process parameters.

At first, the temporal “position” of the ozone exposure was varied, i.e. by performing the anneal during the HfO$_2$ deposition rather than afterwards. Four different conditions were tested, keeping the total exposure time fixed at 5 min: after each cycle of the HfO$_2$ deposition, every 10 cycles, every 25 cycles and after the complete deposition (which was 50 cycles in this experiment). Ozone exposure after each cycle led to increased EOT (Fig. 5.8), which is most likely due to O$_3$-induced oxidation of the Si substrate below the TmSiO layer. Exposure every 10 and 25 cycles did not increase EOT, but provided inferior results in terms of gate current density and interface state density compared to exposure at the end of the deposition (Fig. 5.8a-b).

Having found that the ozone PDA produces the best results when performed at the end of the HfO$_2$ deposition, the duration of the PDA was varied (0–15 min), finding an optimum point around 10 min. Annealing for 10 min provided the lowest gate current density (Fig. 5.9), and on the basis of the strong correlation with the flatband voltage trend it is likely that the effect of the ozone treatment is to reduce negatively charged defects in the TmSiO/HfO$_2$ stack. The duration of the PDA was found to have no effect on EOT and interface state density.

5.2.3 Effect of FGA

All the samples analyzed so far were annealed in forming gas at 400 °C for 15 min at the end of the fabrication process. However, it was found that PDA and FGA
5.3 Implementation of optimized PDA in sub-nm-EOT MOSFETs

Figure 5.9. Gate leakage current density and flatband voltage versus O₃ exposure time for TmSiO/HfO₂ MOS capacitors. (Paper VII)

have a combined effect on gate current and should therefore be co-optimized.

Several types of FGA were tested, varying both time (5–60 min) and temperature (350–400 °C). Higher temperatures (450 °C) and longer times (120 min) were also tested but resulted in strongly degraded EOT and interface state density. Analysis of the different experimental conditions evidenced that PDA in ozone and FGA at 400 °C for at least 15 min are necessary to simultaneously achieve low leakage current density (Fig. 5.10) and low EOT (Fig. 5.11).

FGA was also found to be necessary in order to achieve low interface state density, as exposure to forming gas at 350 °C or 400 °C for 5 min decreases Dᵢₚ by almost one order of magnitude (Fig. 5.12). This is an important result, since it demonstrates that FGA is effective in reducing interface state density even when thulium silicate is used as interfacial layer, or conversely that the interface of TmSiO with Si behaves similarly to that formed by SiOₓ.

5.3 Implementation of optimized PDA in sub-nm-EOT MOSFETs

The optimized PDA (10 min exposure to O₃/O₂/Ar gas flow in the ALD reactor at 350 °C after deposition of HfO₂) was implemented in TmSiO/HfO₂/TiN MOSFETs, using the same thickness of the dielectric layers as the wafer analyzed in Section 5.1 in order to achieve similar EOT and perform a meaningful comparison. FGA was performed at 400 °C for varying cumulative duration in the range 5–60 min. The resulting MOSFET characteristics are presented in Paper VII and the main results are described here in more detail.
5. Enhanced channel mobility at sub-nm EOT

**Figure 5.10.** Gate leakage current density versus duration of the FGA for TmSiO/HfO$_2$ MOS capacitors with different PDA and FGA conditions. (Paper VII)

**Figure 5.11.** EOT versus duration of the FGA for TmSiO/HfO$_2$ MOS capacitors with different PDA and FGA conditions. (Paper VII)
5.3. Implementation of optimized PDA in sub-nm-EOT MOSFETs

Figure 5.12. Interface state density at flatband condition versus duration of the FGA for TmSiO/HfO$_2$ MOS capacitors with different PDA and FGA conditions. (Paper VII)

EOT and channel mobility with and without PDA are compared in Fig. 5.13. It can be observed that the PDA preserved the enhanced mobility of the stack and did not affect EOT on nFETs, while a positive effect may have been induced on pFET EOT. The effect of FGA duration on channel mobility is shown in Fig. 5.14, leading to the conclusion that FGA for at least 15 min is necessary to achieve higher electron mobility than the SiO$_x$/HfO$_2$ reference trend, while hole mobility is almost independent of the FGA duration.

The benefic effect of the combined PDA and FGA treatments on gate leakage current observed in Section 5.2 is confirmed by the MOSFET data, where gate current density is reduced by more than one order of magnitude on both nFETs and pFETs, reaching levels comparable to published trends for state-of-the-art SiO$_x$/HfO$_2$ nFETs (Fig. 5.15).

The effect of FGA on threshold voltages is shown in Fig. 5.16. Before FGA, the EWF of the TmSiO/HfO$_2$/TiN gate stack favors pFET operation, leading to low threshold voltages for pFETs and high threshold voltages for nFETs. However, a negative voltage shift of ~300 mV is observed after 60 min FGA and near-symmetric threshold voltages are achieved. The possibility to tune the threshold voltages by the FGA duration, in addition to the compatibility of the stack with common EWF control techniques (see Section 4.3), indicates that complementary device operation should be achievable in a sub-nm-EOT TmSiO/HfO$_2$/TiN CMOS process.

The FGA-induced voltage shift is also directly visible from the $I_D$-$V_G$ characteristics (Fig. 5.17), which also confirm that both nFETs and pFETs can achieve subthreshold slopes as low as 65–70 mV/dec before FGA. The increased nFET subthreshold slope after FGA can not be due to a degradation of the Si/TmSiO
5. Enhanced channel mobility at sub-nm EOT

Figure 5.13. High-field channel mobility for TmSiO/HfO$_2$/TiN (a) nFETs and (b) pFETs, with and without PDA in ozone. (Paper VII)

Figure 5.14. High-field channel mobility for TmSiO/HfO$_2$/TiN (a) nFETs and (b) pFETs, with PDA in ozone and varying FGA duration.
5.3. Implementation of optimized PDA in sub-nm-EOT MOSFETs

**Figure 5.15.** Gate leakage current density versus EOT for TmSiO/HfO$_2$/TiN nFETs and pFETs, with and without PDA in ozone. (Paper VII)

**Figure 5.16.** Threshold voltage versus FGA duration for TmSiO/HfO$_2$/TiN nFETs and pFETs with PDA in ozone. (Paper VII)
5. Enhanced channel mobility at sub-nm EOT

Figure 5.17. \( I_D-V_G \) characteristics of TmSiO/HfO\(_2\)/TiN nFETs and pFETs, with PDA in ozone and varying FGA duration. (Paper VII)

Interface, since no degradation was observed in nFET mobility (see Fig. 5.14) or in interface state density (see Fig. 5.12, where the nFET interface was probed since MOS capacitors were fabricated on p-type substrates). Given the strong increase in bulk current with increasing FGA time, it is reasonable to assume that long anneals cause formation and/or growth of defects in the source/drain depletion region (due to an insufficient dopant activation anneal), thereby the increased junction leakage contribution to the drain current leads to higher extracted subthreshold slope.

5.4 Reliable device operation

Integration of a novel material in the manufacturing environment can only be possible if it can sustain reliable operation over the expected lifetime (10 years) of CMOS ICs. For a novel MOSFET gate dielectric, this requirement means that reliable device operation must be guaranteed in terms of oxide integrity (measured by TDDB) and stable threshold voltages (measured by BTI).

TDDB was investigated at 125 °C under VRS, which is advantageous in terms of shorter experimental time but provides the same information on breakdown statistics as the conventional CVS [140]. nFETs and pFETs with varying areas were measured at varying ramp rate (Fig. 5.18), for a total of 600 measured devices. A current step of 100 µA in the gate current measured at sense voltage was used as trigger to determine the occurrence of a hard breakdown event. The measured
breakdown voltages (Fig. 5.19a) were transformed into the equivalent times to breakdown (Fig. 5.19b) following the procedure described in [128]. Extrapolation to a reference device area of 0.1 cm² and to a reference failure probability of 0.01% yielded the expected time to breakdown as a function of operating voltage shown in Fig. 5.20. A 10 year lifetime can be guaranteed at a gate voltage of 0.9 V for nFETs and 1.1 V for pFETs, which are reasonable values of operating voltage in future (sub-10 nm) technology nodes.

BTI was measured on nFETs and pFETs at 125 °C under CVS, since this condition allows a straightforward implementation of the stress-relaxation sequence commonly used to separate the recoverable component of the threshold voltage shift from the permanent one, which is the most relevant for reliable long-term operation [141]. A measure-stress-measure (MSM) technique was employed to extrapolate the time dependence of the threshold voltage shift, with stress and relaxation times in the range 0.1–1000 s (Fig. 5.21). Fast (10 µs) measurements were used in order to minimize relaxations in the sensing phase. A 50 mV threshold voltage shift failure criterion was used in extrapolating the expected lifetime as a function of operating voltage, shown in Fig. 5.22. Reliable device operation for 10 years can be achieved at drive voltages of 0.9 V for pFETs and 1.0 V for nFETs.

The results obtained from the TDDB and BTI analyses indicate that integration of TmSiO in HfO₂/TiN MOSFETs is compatible with reliable device operation. Furthermore, the examined devices were not specifically optimized for reliability, therefore improved lifetimes may be obtained by performing additional process development.

Figure 5.18. Gate current measured at stress and sense voltage under VRS in 10 µm × 50 µm nFETs. Application of the stress at varying ramp rate causes a shift in breakdown voltage.
5. Enhanced channel mobility at sub-nm EOT

Figure 5.19. Weibull plot of the (a) breakdown voltage and (b) equivalent time to breakdown distributions measured under VRS on a subset of the tested devices.

Figure 5.20. TDDB lifetime extrapolation on TmSiO/HfO$_2$/TiN nFETs and pFETs. (Paper VII)
5.4. Reliable device operation

Figure 5.21. Drain and gate currents (measured at threshold voltage of the unstressed device) in representative pFETs under different stress voltages.

Figure 5.22. BTI lifetime extrapolation on TmSiO/HfO$_2$/TiN nFETs and pFETs. (Paper VII)
5. Enhanced channel mobility at sub-nm EOT

5.5 Analysis of mobility enhancement

As shown in Sections 5.1 and 5.3, the main advantage in the integration of the TmSiO high-k IL in place of the common chemical oxide lies in the improved EOT-mobility trade-off, since the TmSiO/HfO$_2$ dielectric stack can achieve ~20% higher electron and hole mobility than state-of-the-art SiO$_x$/HfO$_2$ devices.

The mobility enhancement has been investigated by comparing electron mobility in TmSiO/HfO$_2$ devices to published data on scaled SiO$_x$/HfO$_2$ stacks at low, medium and high effective field (Fig. 5.23). The values of inversion charge density ($3 \times 10^{12}$ cm$^{-2}$ and $1 \times 10^{13}$ cm$^{-2}$) used to extract low- and high-field mobility are dictated by availability of literature data, while “medium” field is identified as the region where the mobility peaks. The comparison shows that integration of TmSiO improves electron mobility by 20% at medium and high fields, while little or no improvement is visible at low effective field. A similar comparison was not performed on pFETs due to the lack of literature data at low and medium field, to the best of the author’s knowledge (a 15% enhancement in high-field hole mobility was already demonstrated in Fig. 5.5 in Section 5.1).

In order to clarify the cause of the different performance at low and medium-high effective field, an in-depth analysis of channel mobility at varying temperature and under constant voltage stress has been performed. Analysis of the temperature dependence of channel mobility is a useful technique to identify the scattering mechanisms limiting mobility in the different effective field ranges. To this aim, split-CV electron and hole mobility was measured at varying temperature in the range 250–450 K (Fig. 5.24).

The temperature dependence of mobility at medium effective field is a good indicator of the importance of remote scattering mechanisms (see Section 1.2.2) in high-k/metal gate stacks. As a matter of fact, theoretical calculations have shown that soft-optical phonons in the high-k dielectric are more easily excited than bulk Si phonons, leading to a change in the $T^\alpha$ temperature dependence of phonon-limited mobility from $\alpha = 1.5$ in SiO$_2$-based gate stacks to $\alpha < 1$ in HfO$_2$-based devices [40,142]. Experimental evidence has been provided for $\alpha < 1.5$ in high-k gate stacks, where values in the range 0.8-1.1 have been reported for HfO$_2$-based nFETs with different thickness of the SiO$_x$ interfacial layer and with poly-Si and metal gate electrodes [36,43,142–145], while values in the range 0.6-1.0 have been reported for SiO$_x$/HfO$_2$/poly-Si pFETs [36,144]. The temperature dependence of medium-field mobility for the TmSiO/HfO$_2$/TiN devices is shown in Fig. 5.25. The extracted values of $\alpha$ (0.91 for nFETs and 0.62 for pFETs) are compatible with the ranges reported in literature for HfO$_2$-based devices, indicating that channel mobility in TmSiO/HfO$_2$ stacks is similarly limited by remote phonon scattering in the medium effective field range.

The enhanced mobility can thus be ascribed to the presence of the high-k IL, which presents higher physical thickness at the same EOT of a thinner SiO$_x$ IL. The higher physical thickness of the IL is beneficial in placing the remote scattering sources further away from the channel, since the soft-optical phonon scattering
5.5. Analysis of mobility enhancement

Figure 5.23. Channel mobility in TmSiO/HfO$_2$ nFETs at (a) low, (b) medium and (c) high effective field, compared to literature trends for scaled SiO$_x$/HfO$_2$ devices.
potential has been calculated to decrease exponentially over the interfacial layer thickness [40], as long as the phonon modes in the interfacial layer itself do not significantly interact with carriers in the channel. This is typically the case for materials with reasonably low dielectric constant (up to 10-15), which exhibit low ionic polarizability of the metal-oxygen bonds [40]. An indication that the high-k interfacial layer may indeed reduce soft-optical phonon scattering can be obtained from the temperature sensitivity factor of mobility, defined as \( \frac{d(1/\mu)}{dT} \). Applying Matthiessen’s rule, the factor can be expanded as:

\[
\frac{d(1/\mu)}{dT} = \frac{d(1/\mu_{CS})}{dT} + \frac{d(1/\mu_{ph})}{dT} + \frac{d(1/\mu_{SR})}{dT}
\]

where the first term (< 0) takes into account the temperature dependence of Coulomb scattering, the second term (> 0) takes into account the temperature dependence of phonon scattering and the third term (~ 0) takes into account the temperature dependence of surface roughness scattering [146,147].

High positive values of the sensitivity factor have been observed in SiO\(_x\)/HfO\(_2\)/poly-Si stacks and have been linked to the increased phonon scattering due to the contribution of soft-optical phonons from the high-k dielectric [146]. Lower values (comparable to SiO\(_2\) reference) have, instead, been measured in SiO\(_x\)/HfO\(_2\)/TiN stacks, due to the reduced coupling with gate plasmons, and an increasing trend has been observed with decreasing IL thickness [41]. Fig. 5.26 displays the sensitivity factor of TmSiO/HfO\(_2\)/TiN nFETs and pFETs, where the values obtained in nFETs are extremely close to the values reported for SiO\(_2\)/HfO\(_2\)/TiN stacks with thick (0.9 nm) IL and lower than those with thin (0.7 nm) IL [41].

Figure 5.24. Channel mobility in TmSiO/HfO\(_2\) (a) nFETs and (b) pFETs measured at varying temperature in the range 250–450 K (nFETs have been measured at 250, 275, 300, 325, 350, 375 and 425 K, pFETs have been measured at 250, 300, 350, 400 and 450 K).

5. Enhanced channel mobility at sub-nm EOT
5.5. Analysis of mobility enhancement

Figure 5.25. Channel mobility versus temperature, in logarithmic scale, in TmSiO/HfO$_2$ (a) nFETs and (b) pFETs. The mobility values are taken at $N_{\text{inv}} = 4 \times 10^{12}$ cm$^{-2}$ for nFETs and $3 \times 10^{12}$ cm$^{-2}$ for pFETs, i.e. near the peak.

Figure 5.26. Temperature sensitivity factor of mobility in TmSiO/HfO$_2$ (a) nFETs and (b) pFETs.

result indicates that TmSiO is as effective as a SiO$_x$ layer with the same physical thickness in reducing remote phonon scattering, although it provides a lower IL EOT due to its higher dielectric constant. A similar comparison could not be performed for pFETs due to the lack of reference literature values, to the best of the author’s knowledge.

The different behavior at low effective field, where little or no mobility enhancement was observed, can be explained by the influence of remote Coulomb
scattering from charges in the high-k/metal gate stack. The relative importance of this mechanism can be investigated by stressing the devices under CVS and verifying that a correlation exists between the increased density of trapped charge in the dielectric stack and the degraded device characteristics [42]. Fig. 5.27 shows the measured channel mobility at room temperature after stressing the devices for 0–5000 s. As evidenced in Fig. 5.28, channel mobility is strongly degraded at low-medium field, where remote Coulomb scattering is expected to be relevant, but not at high field. The total increase in trapped charge density ($\Delta N_{OT}$) was estimated from the measured threshold voltage shift, while the density of interface states ($N_{IT}$) was monitored by charge pumping. The increase in bulk charge density can then be obtained as $\Delta N_{BT} = \Delta N_{OT} - \Delta N_{IT}$ (Fig. 5.29), showing that the degradation in interface state density is negligible compared to the increased charge in bulk oxide traps, for both nFETs and pFETs.

Fig. 5.30 shows the strong correlation between the reciprocal of peak mobility and $\Delta N_{OT}$, implying that channel mobility at low-medium field is strongly influenced by remote Coulomb scattering in both nFETs and pFETs. The value of the electron scattering coefficient (2600 Vs/C) is comparable to reported values on SiO$_x$/HfO$_2$ nFETs (2565 Vs/C [42]), indicating a similar influence of remote Coulomb scattering in the two stacks. Hole scattering coefficients for comparable stacks have not been published, to the best of the author’s knowledge. The similar magnitudes of low-field mobility and scattering coefficient for TmSiO/HfO$_2$ and SiO$_x$/HfO$_2$ nFETs implies that the integration of a high-k IL is not beneficial (although it is not detrimental) at low effective field. A possible explanation consists in the fact that, although the increased physical thickness of the IL separates the remote scattering centers from the channel, the higher dielectric constant increases the capacitive coupling of the remote charges to carriers in the channel, as modeled in [38, 65].
5.5. Analysis of mobility enhancement

Figure 5.27. Channel mobility under CVS in representative TmSiO/HfO$_2$ (a) nFET and (b) pFET.

Figure 5.28. Channel mobility at low, medium and high field under CVS in representative TmSiO/HfO$_2$ (a) nFET and (b) pFET.
5. Enhanced channel mobility at sub-nm EOT

Figure 5.29. Interface trap density and increase in bulk and total trapped charge densities under CVS in representative TmSiO/HfO₂ (a) nFET and (b) pFET.

Figure 5.30. Dependence of inverse peak mobility on trapped charge density in representative TmSiO/HfO₂ (a) nFET and (b) pFET.
Conclusions and future outlook

High-k/metal gate technology has become a core enabler of continued CMOS scaling. However, long-term scalability of the widely adopted SiO(N)/HfO$_2$ dielectric stack is hindered by degraded channel mobility, threshold voltage control and device reliability at low EOT. While many works have researched more scalable gate stack designs via integration of novel materials as interfacial layer and/or bulk high-k oxide, this thesis aimed at developing a more structured approach, i.e. finding a manufacturable solution to integrate a high-k interfacial layer with the required properties but maintaining compatibility with both the existing Hf-based technology and plausible future higher-k materials. The main results of this work are summarized here.

Integration of a high-k IL has been identified as the most promising approach to improve the scalability of high-k/metal gate stacks, given the difficulties involved in reducing the contribution of a SiO(N) IL to the total EOT, which are evident from recent literature and were also experimentally verified in the development of a method for chemical oxidation of Si \textit{in situ} in the ALD reactor, presented in Paper I. A successful design process for a high-k IL was identified in the combined research of a material with the desired properties (chiefly a reasonably high dielectric constant and the capability to form a high-quality interface with Si) and of a manufacturable method of forming sub-nm layers in a well-controlled way (comparable to what can be achieved by chemical oxidation and scavenging).

A plausible candidate has been identified in thulium silicate, since it provides the desired material properties and can be grown in the sub-nm regime with the required repeatability via thermal reaction of Tm$_2$O$_3$ with Si. The favorable properties of Tm$_2$O$_3$, in terms of low reactivity toward Si and low hygroscopicity, enabled the development of a straightforward CMOS-compatible process for integration of TmSiO as interfacial layer in a generic gate stack, which was demonstrated in Paper II to be compatible with common gate-last and gate-first integration schemes.

The need to form uniform and conformal layers, especially in view of a future integration in three-dimensional device structures such as FinFETs and nanowires, has motivated the development of a novel ALD process for Tm$_2$O$_3$ deposition, presented in Paper III together with a thorough characterization of the film properties. The main advantage of the developed ALD process, compared to the one that was previously available in literature, consists in the use of a highly reactive
organometallic precursor, which allows the use of H₂O as oxidizing agent, thus avoiding O₃-induced oxidation of the Si substrate. This characteristic, in addition to a good nucleation on HF-last Si surfaces, allows deposition of Tm₂O₃ in direct contact with Si, which is crucial to a well-controlled TmSiO formation.

The TmSiO interfacial layer has then been integrated in a gate-last CMOS process, initially using Tm₂O₃ as bulk high-k oxide to evaluate the properties of the Tm-based material system. The interfacial layer was demonstrated to be competitive toward the common chemical oxide IL and the TmSiO/Tm₂O₃/TiN gate stack provided excellent MOSFET characteristics, in terms of IL EOT, interface state density, subthreshold slopes and channel mobility, as presented in Paper IV.

However, the interfacial layer process was specifically developed for integration in a generic gate stack, without requiring the use of Tm₂O₃ as bulk oxide, therefore the IL was subsequently integrated in a HfO₂/TiN gate stack, which is at present the most relevant high-k/metal system in the microelectronics industry. Comparison of different dielectric stacks (SiOₓ/HfO₂, TmSiO/Tm₂O₃ and TmSiO/HfO₂) has led to the important conclusion that integration of TmSiO does not cause a substantial threshold voltage shift. Indeed, the TmSiO/HfO₂ stack was demonstrated in Paper V to be compatible with common EWF control techniques used in gate-last and gate-first integration schemes, achieving similar EWF tunability to the SiOₓ/HfO₂ dielectric stack.

Integration of the TmSiO IL with HfO₂ as bulk high-k oxide was then demonstrated to be compatible with sub-nm-EOT operation, which is crucial for applicability to future technology nodes. TmSiO/HfO₂/TiN MOSFETs showed excellent device characteristics at EOT as low as 0.65 nm and provided 20% higher electron and hole mobility than state-of-the-art SiOₓ/HfO₂ devices, demonstrating that integration of TmSiO can improve the trade-off between EOT and channel mobility observed at low EOT (Paper VI).

Development of optimized PDA and FGA conditions led to a 10x decrease in gate current density, without affecting EOT, interface state density and channel mobility, and led to the demonstration in Paper VII of reliable device operation over a 10 year lifetime from the point of view of oxide integrity and stable threshold voltages. An in-depth analysis of the obtained mobility enhancement was also performed, leading to the conclusion that the TmSiO IL is as effective as a SiOₓ IL having similar physical thickness in attenuating remote phonon scattering, while providing the advantage of a lower IL EOT.

In summary, TmSiO has been proposed as a candidate high-k interfacial layer for integration in future technology nodes, and especially the TmSiO/HfO₂ dielectric stack has been shown to provide enhanced channel mobility while maintaining compatibility with the requirements of low EOT, controllable threshold voltage and reliable device operation.

However, a few proofs are still needed for TmSiO to be seriously considered as a replacement for the validated chemical oxide technology and should be addressed in future work.

The interfacial layer has been designed for straightforward integration and
evaluated against the requirements of common integration flows, however only long-channel devices have been fabricated in this work. While such devices are excellent test structures for evaluation of novel gate stacks, integration in realistic short-channel devices would be needed to quantify the effects of the TmSiO IL on MOSFET characteristics at short gate length.

Further studies on EOT scalability should also be performed. EOT scaling below 0.6 nm may be achieved by reducing the physical thickness of the TmSiO layer, possibly via optimized silicate formation anneal or interfacial layer scavenging, or by cointegration of a higher-k bulk oxide. The latter approach has been tested by integrating TmSiO with crystalline HfO$_2$, but further experiments are needed to investigate the fabrication of gate stacks with improved long-term scalability by cointegration of a high-k IL and a higher-k bulk oxide.

The concept of a high-k interfacial layer formed by thermal reaction of a deposited oxide with the underlying semiconductor could also be applied to non-Si substrates. Ge channel devices have been proposed as a way to fabricate pFETs (and possibly nFETs) achieving higher performance than conventional (strained) Si devices. Integration of the IL process on Ge should be feasible, as long as a germanate formation reaction can be induced by annealing the Ge/Tm$_2$O$_3$ system and as long as the resulting interfacial layer exhibits all the required properties (especially in terms of thickness control and reproducibility, dielectric constant and quality of the interface with Ge). Some preliminary work has been done on this subject, finding that proper surface cleaning is a critical factor and that the presence of native (sub-)oxide can prevent germanate formation. However, further work is necessary to investigate Tm germanate formation and its suitability as an interfacial layer for Ge-channel MOSFETs.
References


References


References


[56] Based on discussions at the 2013 IEEE Semiconductor Interface Specialists Conference (SISC) and on analysis of the 2013 Edition of the ITRS.


References


J. M. J. Lopes, M. Roeckerath, T. Heeg, E. Rije, J. Schubert, S. Mantl, V. V.


References


References

chemical oxide (Si-O-H) and thermal oxide (SiO₂ or Si-O-N) underlayers.” *Journal of Applied Physics*, volume 92, no. 12, p. 7168, 2002.


References


References

volume 100, no. 1, p. 014504, 2006.

