STELLINGEN

behorende bij het proefschrift

DMOS transistor dope profiling
and
transient enhanced boron diffusion experiments

doors

H.G.A. Huizing

Delft, 16 december 1996.
1. De spanningsafhankelijkheid van "gate overlap"-capaciteiten in MOSFET's kan gebruikt worden om het verloop van een doteringsprofiel langs het oppervlak te meten.  
   Dit proefschrift, hoofdstuk 3.

2. Voor standaard IC-componenten kan via niet-destructieve inverse modellering met behulp van elektrische metingen zelden het volledige doteringsprofiel bepaald worden.

3. Tijdens de eerste seconden van het uitstoken, op 700 °C, van een niet-amorferende silicium zelf-implantatie in een boor-supperrooster, is de diffusiecoefficient van boor vijf ordes hoger dan de evenwichtswaarde.  
   Dit proefschrift, hoofdstuk 4.

4. Het boor dat als gevolg van een silicium-implantatie inactief is geworden, reactivert snel tijdens het uitstoken en be vindt zich derhalve in instabiele clusters.  
   Dit proefschrift, hoofdstuk 4.

5. Het bekende verband tussen de intrinsieke ladersdragerconcentratie en de electriche actieve concentratie van geïmplanteerde en kort uitgestookte boor-profielen wordt niet bevestigd door boor-supperrooster experimenten en berust op toeval.  
   Dit proefschrift, hoofdstuk 4.

6. De doteringsprofielen in toekomstige IC-technologieën zullen in toenemende mate bepaald worden door anomale diffusieverschijnselen. Het is daarom van cruciaal belang om deze verschijnselen te verhelderen.

7. Versnelde diffusie kan in silicium-germanium technologie gebruikt worden om bij relatief lage uitstook-temperaturen (lager dan 700 °C) de fosfor-boor emitter-basis junctie juist op het silicium/silicium-germanium grensvlak te leggen.  

8. De complexiteit van versnelde diffusie als gevolg van het uitstoken van implantatieschade inspireert veel onderzoekers tot het ontwikkelen van ogenschijnlijk nauwkeurige modellen zonder fysiche grondslag.  
9. Anomale diffusie van verontreinigingsatomen hangt af van de kwaliteit van het basismateriaal: hoe schoner dit materiaal is, des te sterker de anomal diffusie zal zijn.

10. Voor een vruchtbare toepassing van Technology Computer Aided Design in de IC-technologie is veel wisselwerking met technologen strikt noodzakelijk.

DMOS transistor dope profiling and transient enhanced boron diffusion experiments
DMOS transistor dope profiling
and
transient enhanced boron diffusion experiments

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus Prof. ir. K.F. Wakker,
in het openbaar te verdedigen ten overstaan van een commissie,
door het College van Dekanen aangewezen,
op maandag 16 december 1996 te 16.00 uur

door

Hendrik Gezienus Albert HUIZING

elektrotechnisch ingenieur
geboren te Emmen
Dit proefschrift is goedgekeurd door de promotoren:

Prof. dr. ir. J.W. Slotboom
Prof. dr. M. Kleefstra

Samenstelling promotiecommissie:

Rector Magnificus (Technische Universiteit Delft)
Prof. dr. ir. J.W. Slotboom (Promotor, Technische Universiteit Delft)
Prof. dr. M. Kleefstra (Promotor, Technische Universiteit Delft)
Prof. dr. C.I.M. Beenakker (Technische Universiteit Delft)
Prof. dr. K. De Meyer (Katholieke Universiteit Leuven, België)
Prof. dr. P. Woerlee (Technische Universiteit Twente)
Prof. dr. N.E.B. Cowern (Aarhus Universitet, Aarhus, Denemarken)
Dr. H. Van der Vlist (Philips Semiconductors Nijmegen)

Huizing, H.G.A.
– With summary in Dutch.

ISBN 90-9010142-X

COPYRIGHT © 1996 by H.G.A. Huizing

All rights reserved.
No part of this material protected by this copyright notice may be reproduced, or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission from the author.

aan Ageeth,
aan mijn ouders,
voor opa.
Contents

1 Introduction ................................. 1
   1.1 Background ............................ 1
   1.2 Definition of the problem ............... 3
       1.2.1 Double diffused MOS transistor .... 3
       1.2.2 Modeling and measuring doping profiles 5
       1.2.3 Efficient optimization ............... 6
   1.3 Outline of the thesis .................... 7

2 Device optimization techniques .......... 11
   2.1 Introduction .......................... 11
   2.2 Response Surface Model method .......... 12
   2.3 Levenberg-Marquardt method ............ 13
   2.4 Combined strategy ....................... 14
   2.5 DC characterization of a lateral DMOS transistor 15
   2.6 Simultaneous optimization of the DMOS and NPN transistor ......................... 18
       2.6.1 Definition of the problem .......... 18
       2.6.2 Calibration of process simulations . 20
       2.6.3 Optimization results ............... 21

3 Determination of doping profiles ......... 29
   3.1 Introduction .......................... 29
   3.2 Existing doping profile measurement techniques .................. 31
       3.2.1 Destructive methods ............... 31
       3.2.2 Nondestructive methods ............ 33
       3.2.3 Inverse methods .................... 35
   3.3 Threshold voltage profiling method ("$V_T$ method") .............. 36
3.3.1 Principle of the method ........................................... 37
3.3.2 Synthetic example .................................................. 39
3.3.3 Limits and restrictions ............................................. 40
3.3.4 DMOS channel profile measurements ............................ 45
3.4 Spline profiling method .............................................. 50
  3.4.1 B-splines and spline functions ............................... 51
  3.4.2 Parameterized doping profiles using spline functions .......... 53
  3.4.3 Implementation .................................................. 54
  3.4.4 Reconstruction of a lateral profile ............................ 56

4 Transient enhanced diffusion ........................................ 63
  4.1 General introduction ............................................... 64
  4.2 Diffusion .......................................................... 65
    4.2.1 Definition of point defects ................................ 66
    4.2.2 Boron diffusion equation and effective diffusivity ....... 67
    4.2.3 The Si-interstitial diffusivity ............................. 71
  4.3 Anomalous diffusion effects ..................................... 74
    4.3.1 Oxidation enhanced diffusion ............................... 74
    4.3.2 Phosphorus enhanced diffusion ............................. 75
    4.3.3 Transient enhanced diffusion ............................... 76
    4.3.4 Dopant clustering and precipitation phenomena ............ 80
    4.3.5 Relation to device behavior ................................ 81
  4.4 Implantation damage annealing and enhanced boron diffusion experiments ......................................................... 82
    4.4.1 How to study damage annealing effects? ................... 83
    4.4.2 Test structures and measurement methods .................. 84
    4.4.3 Initial experiments: clustering and enhanced diffusion .... 87
    4.4.4 Ultrafast injection of silicon interstitials ................. 90
    4.4.5 Boron deactivation .......................................... 99
    4.4.6 Discussion .................................................. 105
    4.4.7 Perspectives ................................................ 107
  4.5 Vertical DMOST profiles .......................................... 109
    4.5.1 Experimental setup ......................................... 110
    4.5.2 Measurement results ........................................ 110
    4.5.3 A modeling exercise ....................................... 113
    4.5.4 Enhanced boron diffusion due to damage annealing after a phosphorus implantation ........................................ 115
## Contents

### 5 Conclusion
- 5.1 Conclusions ................................................. 119
- 5.2 Recommendations for future work ..................... 122

### References ..................................................... 125

### List of symbols and abbreviations ......................... 143

### Summary ....................................................... 147

### Samenvatting .................................................. 151

### Biography ..................................................... 155

### Acknowledgements ........................................... 157
Chapter 1

Introduction

1.1 Background

Since the invention of the bipolar transistor [6] and, later, the emergence of the feasible MOSFET and the integrated circuit (IC), silicon IC technology has evolved dramatically. Today the impact of ICs is omnipresent. To enable this evolution a thorough knowledge of process and device physics is required. With the increasing complexity of the technologies, Technology Computer Aided Design (TCAD) has become an important tool to enhance insight and to shorten the ‘time to market’.

For instance, circuit simulators were developed [110] to assist IC designers, and they are currently an indispensable tool for IC design. These simulators are based on accurate — compact — analytical models for IC technology components. Such models are computationally very efficient and can predict the IC performance within reasonable time, even for large and complex ICs. An important issue in IC technology arises from parasitic capacitances and inductances in IC layouts due to long interconnects and multiple interconnect layers. To include these parasitics in the circuit simulations, layout extraction software has become available [7].

Device simulators simulate electrical characteristics of devices [56, 142, 150]. Based on the numerical solution of Poisson’s equation and the continuity equations for charge carriers, these simulators are able to give insight into problems which cannot be solved analytically; it is, for instance, possible to use arbitrary doping profiles. Device simulators are now considered to be rather accurate, as a lot of effort has been made to determine important physical parameters, such as the charge carrier lifetimes and mobilities. In emerging technologies, device simulators will play an increasingly important role in the optimization of device designs.
Process simulators were developed to simulate the device structure, serving as input for device simulators [2, 177]. In addition to the solution of Fick's law for dilute diffusing species, these simulators incorporate models to describe oxidation, epitaxial growth, ion implantation, deposition of—possibly doped—materials and even lithography. Due to the large amount and complexity of the processing phenomena, accurate prediction of, e.g., doping profiles is still difficult.

To study the implications of technological changes on device or circuit behavior in a structured way, the TCAD software has been extended by optimization tools such as the Response Surface Model method which is based on a Design Of Experiments strategy. For this purpose, software has become available to link simulators automatically [16]. In this setup, the technology is optimized in terms of device and circuit performance. Figure 1.1 is an illustration of linked simulators.

![Diagram of simulator chain](image)

Figure 1.1: Simulator chain for optimization of IC technology and circuit performance. On the right icons symbolize the result at each stage. (e.g. transistor, IV-characteristics, parameter extraction and pass filter).

Although automated linking can, in principle, be done up to any level of abstraction, it must be emphasized that only a few links are practical, in view of computation time and error accumulation.

Prior to device optimization, the chain shown in figure 1.1 may be used to calibrate the physical models underlying the process and device simulators by using
measurement results from dedicated test structures. A simple example is the determination of doping diffusion coefficients by fitting simulated doping profiles to measured ones.

Finally, the chain can be used for the Inverse Modeling (IM) of devices [30, 124]. A practical example is the extraction of parameterized doping profiles from MOS-CV measurements, see, e.g., Osse [119] and Ouwerling [120].

The usefulness of TCAD for technology optimization obviously depends on the accuracy of its building blocks. While device simulations are rather accurate, their results depend strongly on a correct description of the geometry and doping profiles supplied by process simulations. However, accurate process simulations are often still difficult due to the previously mentioned large set of phenomena to be modelled, and the lack of accurate two-dimensional doping profiling methods to verify such models. Moreover, when 2D simulations are required, process simulations can be rather time consuming which reduces their usefulness for global device optimizations considerably.

1.2 Definition of the problem

The objective of the work presented in this thesis was the simultaneous optimization of a double diffused power MOS (DMOS) and the corresponding bipolar NPN transistor using TCAD tools. A reliable and straightforward implementation was hampered by uncertainties in the modeling of the 2D doping profiles and by unacceptably long computation times, both being commonly encountered problems. These problems are addressed in this thesis and will be discussed in the following subsections after a more detailed description of the DMOST and NPN.

1.2.1 Double diffused MOS transistor

The integration of MOS power devices into monolithic IC technologies is important to create smaller electronic power systems and ICs with high current-driving capabilities. It allows the integration of relatively expensive power devices and control/protection circuitry on one chip, which considerably reduces the cost, size and complexity of electronic power systems. These "smart power" devices therefore find applications in, e.g., automotive and battery charging electronics and have recently received much attention.

The DMOS power transistor is particularly suited for monolithic integration with
bipolar transistors, as it has essentially a planar bipolar transistor structure (see figure 1.2).

![Diagram of a vertical DMOS transistor and an NPN created with the same, double diffused, profiles.]

Figure 1.2: A vertical DMOS transistor (VDMOST) and an NPN created with the same, double diffused, profiles.

In DMOS devices, the channel (back gate region) and the source are created by self-aligned diffusions using the gate as mask. Originally, DMOS transistors were proposed as a short-channel, high-frequency alternative to NMOS transistors [59, 147, 163]. As the channel length is determined by lateral diffusion instead of lithography, relatively small channel lengths and fast MOS transistors could be made. Lithography, however, has made considerable progress towards smaller dimensions and there is currently no need to apply the DMOS structure, which has several disadvantages such as a lower package density, a higher spread of the threshold voltage and more complex processing. For power applications, however, the DMOS is still a very attractive device because the channel region is relatively in-
sensitive to the drain voltage, due to an independently adjustable low drain doping level. This allows a high drain source voltage. The threshold voltage is determined by the maximum net doping level in the channel region which depends on the lateral out-diffusion of both channel and source doping profile and is, therefore, sensitive to process variations.

The DMOS profiles are also used to define an NPN transistor, which is used in the control circuitry. This saves two masks and additional processing steps, which reduces the cost and increases the yield of the process.

Obviously, the electrical characteristics of both devices are coupled and they cannot be optimized independently. Also, the simultaneous optimization of both devices is essentially a two-dimensional problem, as the DMOS is created by laterally out-diffusing doping profiles. The optimization problem is complex and, therefore, TCAD tools must be used to determine the trade-offs and to assist the device optimization without performing expensive wafer processing.

Although TCAD tools are today available in many laboratories, numerical device optimization is still not a common step in technology optimization. The main reasons are the difficulty of modeling the process accurately and the long computation times involved. While studying the DMOST and NPN, we soon encountered these problems, which hampered a straightforward, reliable, TCAD optimization. We treat them in more detail in the following subsections.

1.2.2 Modeling and measuring doping profiles

As already mentioned, the most critical part in the optimization chain of figure 1.1 is an accurate prediction of the two-dimensional doping profiles, i.e. the process simulation. Process simulations are complex and their predictability is often limited, although they give useful results for vertical profiles for commonly met situations. As an illustration, in figure 1.3 secondary ion mass spectrometry (SIMS) measurements are shown of the emitter and base doping profiles of the NPN transistor together with default TSUPREM4 [166] simulations in figure 1.3.

Although the important vertical junction depth is determined correctly, the phosphorus profile is, qualitatively, not accurately simulated. Without further measures, straightforward application of TSUPREM4 for accurate device optimizations seems questionable.

The physical models of process simulators must be verified and calibrated against measurements. There are accurate and widespread measurements techniques for
1. Introduction

Figure 1.3: SIMS measurements and default TSUPREM4 simulations of the emitter and base profiles in figure 1.2. The phosphorus dose which was found in the measurement was lower than expected.

the measurement of vertical doping profiles, such as secondary ion mass spectrometry and MOS-CV techniques, which allow a straightforward 1D calibration of these simulators. However, for a verification of the physical models and the simulation results in two dimensions, 2D measurements are preferable. Many such methods have been proposed in the last decade, but they have not yet matured to a level where they give results that are as accurate as the 1D analysis techniques. Moreover, most 2D methods require difficult sample preparation and/or data interpretation, and are not used routinely. Thus, no extensive database of accurate 2D impurity profiles has yet been established. This impedes a straightforward calibration of 2D process simulation models and strongly limits the accuracy of TCAD device optimizations.

1.2.3 Efficient optimization

The calibration of simulators and the final optimization require much computer time, as 2D simulations, in general, are time consuming. Moreover, in practice, it is not always clear in advance which process parameters will turn out to be important for the optimization. This results in an initially large set of parameters
1.3. Outline of the thesis

which also increases the total computation time. To end up with overall acceptable computation times, an efficient and robust optimizer, or optimization strategy, is needed capable of handling an initially large set of parameters of which some may turn out to be insensitive.

1.3 Outline of the thesis

In this thesis, work on three subjects is presented. The work is a result of a survey of the modeling, measuring and optimization problems discussed in the previous section.

An efficient optimization strategy

A new optimization strategy is proposed in chapter 2 to control overall computation times in TCAD optimizations. It is based on a combination of two existing methods, i.e., the Response Surface Model and Levenberg-Marquardt methods. It is argued that these methods can be used complementarily, combining their strengths while avoiding their weaknesses. The combination is a potential powerful tool to find precise optima with a maximum of information from a limited number of experiments.

We first discuss both methods separately and point out their respective advantages and disadvantages. Next, we introduce the combined strategy and provide two examples.

Determination of doping profiles

In chapter 3, we present two novel nondestructive measurement techniques of the lateral doping profiles along the interface in the DMOST.

The first measurement method is a direct lateral profiling method based on a threshold voltage formulation. It is essentially a 1D method and can be used for a fast inspection of a lateral profile. It uses $CV$ measurements only and is, therefore, particularly suited for monitoring in-line processes. A "synthetic" example shows that the method gives good results for profiles that are comparable with the back gate profile of the DMOST. A discussion on the limits and restrictions of this method is provided. Measurement results for DMOS channel profiles show the excellent sensitivity for profile variations.
Next, a second profiling technique, which utilizes the concept of Inverse Modeling (IM), is presented. Unlike the first method, IM does not require assumptions to directly calculate the doping profile from the measurements. During IM, an optimizer iteratively improves a parameterized doping profile by minimizing the difference between simulated and measured device characteristics. To describe the doping profiles, analytical models or a process simulator may be used. However, such models lack global flexibility as unmodeled features cannot be resolved. Therefore, after an idea by Khalil [79], spline functions are used to describe the 2D doping profiles. A spline function can mimic any continuous shape with a limited number of parameters, and is, therefore, very attractive to use as a parameterized model that is capable of describing arbitrary doping profiles.

The spline calculation routines are implemented in a computer program which communicates with numerical simulators through the standardized TCAD files of the widespread and readily available TMA (Technology Modeling Associates) software tools. The program replaces the doping values by the spline function values in predefined parts of the structure and leaves the remaining part unchanged.

The method can, in principle, be applied to 2D profile extraction, provided that the measurement data comprises 2D profile information. Our test structures, however, were initially designed for the lateral profiling method, and they were not suited to a full 2D profile extraction. Yet, in order to show the feasibility and flexibility of the spline profiling method, we show the extraction of the lateral profile under the gate of a DMOST without a source, i.e. a gated diode.

**Transient enhanced diffusion**

In chapter 4, experiments on enhanced diffusion phenomena during ion implantation damage annealing are presented.

This work was initially motivated by the observation of transient enhanced diffusion of the doping profiles in the DMOS shown in figure 1.2. Also, enhanced diffusion phenomena play a crucial role in the doping profiles and device characteristics in other advanced MOS and bipolar device technologies. These phenomena are still not well understood, and they are currently one of the features that should most urgently be included in process simulators to enable the modeling of forthcoming technologies. Finally, the work was stimulated by the ability to grow high-quality intrinsic silicon epitaxial layers with ASM-Epsilon chemical vapor deposition apparatus. The low density of defects in this material, in combination with the possibility to incorporate narrow boron-doped layers, allows easy detec-
tion of enhanced diffusion phenomena by, e.g., SIMS analysis.

![Graph showing enhancement of boron diffusion coefficient during damage annealing after a medium dose ion implantation.](image)

Figure 1.4: Enhancement of the boron diffusion coefficient during damage annealing after a medium dose ion implantation.

The temperature of 700 °C in our experiments is such that the results are especially interesting for new IC technologies, such as those based on thin silicon-germanium layers grown at 700 °C. At these temperatures transient enhanced diffusion effects are more pronounced than at higher temperatures. After an introduction to anomalous diffusion, we report an ultrafast diffusion enhancement preceding a transient phase already known in the literature. Figure 1.4 depicts the enhancement of the boron diffusion coefficient during damage annealing after ion implantation. Here, the implant dose is considered low enough to avoid the formation of dislocation loops [76]. At 1000 °C a brief period of enhanced diffusion has been observed and ascribed to dissolving point defect clusters. At 700 °C this transient is both longer and stronger [151].

Further, boron deactivation due to silicon self-implantations and deactivation during implant damage annealing has been studied. Finally, we analyze the profile evolution of the DMOS profiles, in view of damage annealing effects.
1. Introduction
Chapter 2

Device optimization techniques

In this chapter, we propose to use a combination of two optimization methods, the Response Surface Model method and the Levenberg-Marquardt method, for various optimization issues in IC technology. It is argued that both optimization techniques can be used in a complementary way, combining their strengths while avoiding their weaknesses. This allows the setting up of efficient optimizations that yield precise optima with a limited set of numerical experiments.

After an introduction, we briefly introduce the Response Surface Model and the Levenberg-Marquardt method and discuss their advantages and disadvantages. Next, we introduce the combined strategy and conclude with two examples.

2.1 Introduction

Over the years, several gradient-based optimization techniques (e.g., Levenberg-Marquardt (LM)) have been studied and applied to optimization issues in IC technology (e.g., Doganis [33], Sakurai [137], Ouwerling [123]). More recently, Response Surface Model (RSM) methods have been demonstrated as an alternative, see, e.g., Cartuyvels [16].

In the RSM method, a series of predetermined —computer— experiments are carried out, covering the whole space of allowed parameter settings. Next, analytical models are fitted to the calculated results. These approximating models are a computational fast alternative for the numerical models and can be used for optimization purposes. The optimum thus found, however, may deviate from the actual optimum due to a possible poor predictability of the RSMs.

The LM method proceeds towards an optimum by conducting the actual nu-
merical experiments and does thus not rely on intermediate approximating models. A major drawback, however, is that it gives only a local view of the optimization problem because the solution is based on the last few experiments.

The combination of the global approach of the RSM method with the local accuracy of the LM method is a potentially powerful tool to find precise optima with a maximum of information from a minimal number of experiments [70]. In this setup, the RSM method collects *a priori* information for an efficient optimization by the LM method.

### 2.2 Response Surface Model method

The RSM method is based on a Design Of Experiments (DOE) strategy for planning experiments which cover the entire relevant parameter space. Box [12] gives a good and practical introduction, including examples, in the DOE/RSM method.

In the context of numerical device optimization, “experiment” refers to a computer experiment, e.g., a device simulation, and the parameters are, for instance, oxide thicknesses and implantation doses. The results of these “experiments”, e.g., current gains or cut-off frequencies, are called responses. The surface, spanned by a response over the parameter space, is referred to as response surface.

Using these “experiments”, an empirical model (*response surface model*) is fitted to each response surface. Hence, these approximating RSMs may be used instead of the numerical models themselves for very fast *ad hoc* ‘simulations’, statistical studies or parameter optimization. Obviously, the results rely on the accuracy of the RSM.

Provided there is a good fit of the RSM to the “experiments”, this method has important advantages:

- optimizations are very fast and can easily be repeated for different targets and weights;
- insight in global behavior;
- easy determination of relevant parameters;
- detection of correlated parameters;
- detection of local optima;
- identification of the global optimum.
2.3. Levenberg-Marquardt method

The main disadvantages are:

- the need for a smooth response in order to be able to fit an analytical model;
- and, related, an error due to a possible poor fit of the RSM to the experiments.

2.3 Levenberg-Marquardt method

The gradient-based direct optimization Levenberg-Marquardt (LM) method was introduced by Marquardt [102] after an idea by Levenberg [91]. Although various optimization methods exist, the LM method has become the standard of nonlinear least-squares routines. It is based on the fast, but unstable, Newton method and the robust, but often inefficient, steepest descent method. It minimizes the least-squares error:

$$\xi(\vec{p}) = \frac{1}{N} \sum_{j=1}^{N} w_j^2 \left[ y_j^{\text{meas}} - y_j^{\text{calc}} \right]^2$$

(2.1)

where $N$ is the number of data points, $\vec{p}$ the parameter vector, $w_j$ a weight factor, $y_j^{\text{meas}}$ the measured data and $y_j^{\text{calc}}$ the calculated data. Often $w_j$ is chosen to be the inverse of $y_j^{\text{meas}}$ to take into account relative errors. In the case of noisy data, the standard deviation per data point, if known, may be used. ($\chi^2$-fitting).

For instance, in a device optimization, the error sum may be constructed from target device parameters such as threshold voltages, current gains, or breakdown voltages which are extracted from the simulated characteristics. In general, such parameters appear in compact analytical models, so, alternatively, the error sum may be constructed using electrical characteristics which are calculated with these models [68].

We note the following advantages of the LM method in relation to numerical optimization of devices:

- provided a good initial guess, the method efficiently proceeds towards the optimum;
- the solution is as accurate as the numerical models allow, in contrast to the RSM method which relies on intermediate approximating models.
However, the LM method also has typical disadvantages:

- it can be trapped into a local optimum;

- only a local view on parameter sensitivity and correlation can be obtained;

- insensitive input parameters take unphysical values and increase the total computation time needlessly;

- if the targets or weights are changed, a full new iteration loop has to be initiated.

For an efficient LM optimization, only sensitive and independent parameters must be considered. Moreover, a good starting point is required and the parameters must be properly constrained to avoid trapping in a possible nearby local optimum. A priori knowledge therefore is important to ensure an efficient LM iteration.

### 2.4 Combined strategy

The disadvantages of the LM and RSM methods in practice often means that most time is spent in the conditioning of the optimization problem, in order to arrive at a realistic optimum. The combination of both methods can relieve us of this problem.

First, we use the RSM method to get a global view on the parameters, allowing us to screen out less sensitive parameters. Additionally, the RSM indicates trends in parameter space leading to an optimal solution. Next, we find the optimum in the RSM and supply this as a starting guess to the LM algorithm for a fine-tuning of the parameter values. Since the RSM is an approximation of the unknown response, the optimum based on the RSM will be close to the actual optimum. Hence, it serves as a good starting point for the LM technique.

In collaboration with IMEC in Belgium, the combined strategy has been implemented by A. Sadeghzadeh [136] by writing and testing an interface between the RSM package NORMAN/DEBORA [10,16] and the general data processor PROFILE [121], which features an LM optimization routine. It is organized as follows. PROFILE is the driver program during the nonlinear optimization. The forward model in figure 2.1 is already defined in the NORMAN input file and PROFILE therefore runs NORMAN to invoke this model. On invocation, NORMAN reads the current parameter values from file and runs the forward model once. The simulation results during the nonlinear optimization are saved by NORMAN and are added to
2.5 DC characterization of a lateral DMOS transistor

As a fast and simple application, we show the calibration of a compact MOS model on $I_D(V_G)$ characteristics of a lateral double diffused MOS-transistor (LDMOST).

Figure 2.1: Illustration of general nonlinear optimization. A priori information is needed and can be obtained in a structured and efficient way by using Response Surface Modeling. As an example of a —nonlinear— forward model linked process and device simulators are shown.

the existing set of data. The data generated in the LM loop can thus be re-used for further fine-tuning of the RSM model in the neighbourhood of the optimum. To assist NORMAN users, the PROFILE input files can be automatically generated by invoking NORMAN with a proper argument.

2.5 DC characterization of a lateral DMOS transistor

As a fast and simple application, we show the calibration of a compact MOS model on $I_D(V_G)$ characteristics of a lateral double diffused MOS-transistor (LDMOST). (This is one of the test devices to be treated in section 3.3). As the model evaluation is fast, the need for an efficient optimizer is not urgent in this particular example, but it serves well as a demo. Figure 2.2 is an illustration of the measured LDMOST. To model the low drain bias $I_D(V_G)$-characteristics, the LDMOST can be thought
of as being an intrinsic MOS with a gate bias dependent drain series resistance. The following compact model is capable of describing these characteristics (see, e.g., De Graaff and Klaassen [53] for compact models):

\[
I_D = \frac{\beta(V_G - V_T)V_D}{1 + \theta_A(V_G - V_T) + \theta_B(\sqrt{-V_B + 2\phi_F} - \sqrt{2\phi_F})}
\]  

(2.2)

where:

\[
V_T = V_{t0} + \Gamma(\sqrt{-V_B + 2\phi_F} - \sqrt{2\phi_F})
\]  

(2.3)

\[
\Gamma = \begin{cases} 
\gamma & -1 < V_B < 0 \\
\gamma_I & V_B < -1 
\end{cases}
\]  

(2.4)

\[
\theta_A = \theta_{A0} + \beta(R_{t0} + R_{acc} || R_b)
\]  

(2.5)

\[
R_{acc} \approx \frac{1}{\mu_{0}e_{0}z} \frac{L_{acc}}{W} V_G
\]  

(2.6)

where the bold-faced parameters (\(\beta, \theta_B, V_{t0}, \gamma, \gamma_I, \theta_{A0}, R_{t0}, \) and, \(R_b\)) are candidates for the fitting procedure. \(\beta\) is the gain factor, \(\theta_A\) and \(\theta_B\) are the mobility reduction factors due to normal and lateral electric fields in the channel, \(\Gamma\) is the
Simultaneous optimization of . . .

body effect, described by two coefficients to model the nonuniform channel doping profile, $R_{t0}$ is a constant series drain-source resistance, and $V_{t0}$, finally, is the zero back bias threshold voltage. For conceptual and computational convenience, we modeled the gate voltage dependent drain resistance through the mobility reduction parameters, which is allowed in the studied bias situation. From equation 2.5, it is clear in advance that it is not possible to distinguish between $\theta_{A0}$ and $R_{t0}$ and we therefore lump these parameters by presenting one value for $\theta_{A0}/\beta + R_{t0}$.

Table 2.1: Extracted compact model parameters of the fits shown in figure 2.3 for the RSM method and, consecutively, the LM method. The parameters which are marked ‘insensitive’ turned out to have no effect on the fit and were not considered in the LM procedure.

<table>
<thead>
<tr>
<th></th>
<th>unit</th>
<th>RSM</th>
<th>+LM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>$A/V^2$</td>
<td>0.584</td>
<td>0.603</td>
</tr>
<tr>
<td>$V_{t0}$</td>
<td>$V$</td>
<td>2.48</td>
<td>2.46</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>$V^{1/2}$</td>
<td>2.41</td>
<td>2.21</td>
</tr>
<tr>
<td>$\gamma_i$</td>
<td>$V^{1/2}$</td>
<td>insensitive</td>
<td>-</td>
</tr>
<tr>
<td>$\theta_B$</td>
<td>$V^{-1/2}$</td>
<td>insensitive</td>
<td>-</td>
</tr>
<tr>
<td>$\frac{\theta_{A0}}{\beta} + R_{t0}$</td>
<td>$\Omega$</td>
<td>2.58</td>
<td>2.44</td>
</tr>
<tr>
<td>$R_b$</td>
<td>$\Omega$</td>
<td>1.79</td>
<td>1.49</td>
</tr>
</tbody>
</table>

Using the RSM method, we studied the linear error of each $I(V)$ curve as a function of all the parameters in the MOS-model. The results indicated that parameters $\gamma_i$ and $\theta_B$ have little influence; the remaining parameters were optimized using the response surfaces and the optimum was supplied as a starting point to the LM routine. Table 2.1 shows the results for the parameters, and figure 2.3 shows the fits found by the RSM method and, consecutively, the LM method.

The RSM method performed 143 experiments to collect the a priori information. Using the RSM results to set up an LM optimization, the LM routine only needed 5 iterations and 20 model evaluations to converge to the optimum. When using only the LM routine without the RSM method, no indication about the parameter sensitivity and the location of the minimum is available, and the rate of convergence very much depends on the initial guess. When using all seven parameters and random initial guesses, it took at least 12 iterations and 50 model evaluations in which $\gamma_i$ and $\theta_B$ obtained random values due to their insensitivity.
2. Device optimization techniques

Figure 2.3: Results of the calibration: one RSM iteration gave the most critical parameters and an indication of the optimal fit which is used as a starting point for the LM routine. Only drain currents higher than 5 mA were used in the fitting procedure.

2.6 Simultaneous optimization of the DMOS and NPN transistor

In order to illustrate the feasibility of a complex device optimization we present a simultaneous optimization of the DMOS and corresponding NPN, being the first objective of the work in this thesis. The example may be considered part of the optimization of the entire process focusing on the DMOS and NPN devices.

We first give the definition of the optimization problem, followed by a calibration of the simulators using available measurement results. Finally, we present the optimization results.

2.6.1 Definition of the problem

In table 2.2 we summarized the device parameters which we intend to optimize, together with the process parameters which we initially consider.

The DMOST device parameters are the threshold voltage $V_T$ and the specific
Table 2.2: Process and device parameters which are considered for the simultaneous optimization of the VDMOST and the NPN. The nominal values of the parameters are currently in use c.q. typically measured.

<table>
<thead>
<tr>
<th>process param.</th>
<th>units</th>
<th>nominal value</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_{epi}$</td>
<td>cm$^{-3}$</td>
<td>3.6$\times10^{15}$</td>
<td>epi-layer concentration</td>
</tr>
<tr>
<td>$t_{epi}$</td>
<td>$\mu$m</td>
<td>9</td>
<td>epi-layer thickness</td>
</tr>
<tr>
<td>$T_{ox}$</td>
<td>min</td>
<td>67 ((\equiv) 70 nm)</td>
<td>gate oxidation time</td>
</tr>
<tr>
<td>$L_G$</td>
<td>$\mu$m</td>
<td>10</td>
<td>drawn gate length</td>
</tr>
<tr>
<td>$\Phi_{SP}$</td>
<td>cm$^{-2}$</td>
<td>7.5$\times10^{13}$</td>
<td>SP implant dose</td>
</tr>
<tr>
<td>$T_{SP}$</td>
<td>min</td>
<td>480</td>
<td>SP anneal time</td>
</tr>
<tr>
<td>$\Phi_{SN}$</td>
<td>cm$^{-2}$</td>
<td>5.0$\times10^{15}$</td>
<td>SN implant dose</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>device param.</th>
<th>units</th>
<th>nominal value</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_T$</td>
<td>V</td>
<td>(\pm) 2.5</td>
<td>VDMOST threshold voltage</td>
</tr>
<tr>
<td>$R_{on}^{sp}$</td>
<td>$\Omega$ mm$^2$</td>
<td>0.3-0.35</td>
<td>specific on-resistance</td>
</tr>
<tr>
<td>$BV_{DS}$</td>
<td>V</td>
<td>80-90</td>
<td>breakdown voltage</td>
</tr>
<tr>
<td>$h_{FE}$</td>
<td>-</td>
<td>80-100</td>
<td>NPN dc current gain</td>
</tr>
<tr>
<td>$f_T$</td>
<td>MHz</td>
<td>300-350</td>
<td>cut-off frequency</td>
</tr>
<tr>
<td>$U_A$</td>
<td>V</td>
<td>400-500</td>
<td>Early voltage</td>
</tr>
</tbody>
</table>

on-resistance of the device $R_{on}^{sp}$. The NPN device parameters are the dc-current gain $h_{FE}$, the cut-off frequency $f_T$ and the Early voltage $U_A$. The drain-source breakdown voltage $BV_{DS}$ is not an optimization parameter but indirectly controlled by the other device parameters such as the base doping. Obviously, we will verify $BV_{DS}$ after the optimization is completed.

The drawn gate length $L_G$ and the gate oxidation time $T_{ox}$ affect the DMOST only. The nominal value of $T_{ox}$ is 67 minutes, which corresponds to an oxide thickness of 70 nm. The other process parameters are also of importance to the NPN: the SN-dose $\Phi_{SN}$ ("Shallow N", used to create the n-type source, drain and emitter regions), SP dose $\Phi_{SP}$ ("Shallow P", used to create p-type back gate and base regions), SP anneal time $T_{SP}$, the epi-layer concentration $N_{epi}$, and, finally, the epi-layer thickness $t_{epi}$. Some, such as $T_{ox}$ or $N_{epi}$, will also affect the behavior of other devices in the process, and can therefore not be optimized independently. They are, however, initially included to obtain an overall view of the optimization problem. In the final stage, we tune the process parameters which are only used for the VDMOST and NPN, i.e. $L_G$, $\Phi_{SP}$, and $T_{SP}$. 
We used the efficient 1D process simulator SUPREM3 [165] to calculate the vertical doping profiles. The required two-dimensional profiles are created by extending the simulated vertical SN and SP doping profiles laterally by using a complementary error function. To enable this, the ratio of the lateral and vertical characteristic length of the profiles must be known. This is determined in the next subsection. The device simulator MEDICI [164] is used to calculate the target device parameters for both the VDMOST and NPN.

The threshold voltage \(V_T\) is defined by the intercept of the voltage axis and the tangent line at the steepest part of the \(I_D(V_G)\)-curve with \(V_D=0.1\) V. The specific on-resistance \(R_{on}^p\) is calculated using \(V_D = 0.1\) V and \(V_G = 13\) V. The breakdown voltage \(BVDS\) is calculated with all terminals at zero bias, except the drain, and is determined by the drain voltage at which the maximal impact ionization integral equals one [159, 164].

The current gain \(h_{FE}\) and cut-off frequency \(f_T\) are determined by their highest value. They are calculated by sweeping the emitter bias from zero to \(-1\) V while keeping the base and collector at zero bias.

### 2.6.2 Calibration of process simulations

Before launching the device optimization, we first need to calibrate the process simulations on existing SN and SP doping profiles by using the following available measurements of the nominal devices:

- Doping profile measurement of the SP profile (by secondary ion mass spectrometry analysis). The main drive for SP is an anneal of 480 minutes at 1100 °C;

- The backbias dependence of \(V_T\) of the LDMOST structure which allows to probe the DMOST channel profile into depth. Measurements were done on wafers from two test batches, labeled ECI389 and ECI558.

**Adapted model parameters**

It appeared from profile measurements, that the boron diffusion was less than simulated by SUPREM3, and we therefore adapted the boron diffusion coefficient \(D_B\). We did not change other SUPREM3 model parameters. The 1D simulations were used to set up a 2D profile. The required ratios of the lateral and vertical characteristic lengths (\(\sigma_y^{SN}/\sigma_x^{SN}\) and \(\sigma_y^{SP}/\sigma_x^{SP}\), \(x\) being the vertical direction) were determined by fitting the \(V_T(V_B)\) characteristics of the nominal LDMOST. We used
the LDMOST for this purpose rather than the VDMOST as it has the required separate back gate contact.

Calibration results

The results of the calibration are presented in figures 2.4 and 2.5 and in table 2.3. In figure 2.4a the SIMS measurement and simulation results are shown. To obtain a good fit, a smaller boron diffusion coefficient had to be taken. The SIMS was taken from the test batch described in chapter 4 and had an SP dose of \(7 \times 10^{13}\) cm\(^{-2}\) instead of the default default dose of \(7.5 \times 10^{13}\) cm\(^{-2}\). The lateral characteristic length of the source and back gate profile were determined by using the \(V_s(I_B)\) measurements for \(V_{SB}=0, 2, 4\) and \(6\) V of the ECI389 test wafer. The results after calibration are shown in figure 2.4b. More points are shown for clarity. Also, measurements and simulations of the wafers of batch ECI558 are included. In addition to the nominal devices, in batch ECI558, two more types were available, differing only by a 10% variation of the SP implant dose \(\Phi_{SP}\), which enables a cross-check of the extracted parameters.

A further check of the extracted profiles is shown in figure 2.5. The extracted profiles were used to simulate the characteristics of an NPN transistor with an emitter area of \(24 \times 6\) \(\mu\)m\(^2\). Shown are the simulations and measurements of the Gummel plot and the \(f_T\) characteristics. The \(f_T\) measurements were taken from a different wafer of the ECI558 batch and supplied by Ir. M. Swanenberg from Philips Consumer IC division, Nijmegen, the Netherlands. Of the devices with 10% SP variations, no \(f_T\) measurements were available.

The sheet resistances as well as the device characteristics of batch ECI558, all calculated with the same ratio \(\sigma_y^{SN}/\sigma_x^{SN}\) and \(\sigma_y^{SP}/\sigma_x^{SP}\), are given in table 2.3. Considering the process spread, the simulation results are satisfactory.

2.6.3 Optimization results

After the calibration, NORMAN is used to perform a series of numerical experiments. The considered process parameters have already been given in table 2.2 and repeated in table 2.4 where the experimental ranges are included. Using these experiments, NORMAN calculates RSMs which are then used to determine the most sensitive process parameters, possible trade-offs and an initial optimum. A second
Figure 2.4: Measurements and simulation results after calibration of the doping profiles.  
a) Boron doping profile; to determine the boron diffusion coefficient.  
b) \( V_T(V_B) \) characteristics; to determine the lateral characteristic lengths of the source and back gate profile.
Figure 2.5: Verification of the calibrated doping profiles for the $W \times L = 24 \times 6 \, \mu m^2$ NPN transistor. 

- **a)** Gummel plot.
- **b)** $f_T$ simulations for various collector base voltages.
2. Device optimization techniques

Table 2.3: VDMOST and NPN Device charactersics after calibration of the 2D doping profiles for varying SP doses (batch ECI558).

<table>
<thead>
<tr>
<th>model parameter</th>
<th>units</th>
<th>default value</th>
<th>calibrated value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_B$ (1100 °C)</td>
<td>cm$^2$s$^{-1}$</td>
<td>5.4×10$^{-10}$</td>
<td>4.0×10$^{-10}$</td>
</tr>
<tr>
<td>$\sigma_{y}^{SN}/\sigma_{x}^{SN}$</td>
<td>-</td>
<td>-</td>
<td>0.707</td>
</tr>
<tr>
<td>$\sigma_{y}^{SP}/\sigma_{x}^{SP}$</td>
<td>-</td>
<td>-</td>
<td>0.959</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>0.9·$\Phi_{SP}$</th>
<th>$\Phi_{SP}$</th>
<th>1.1·$\Phi_{SP}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>device parameter</td>
<td>units</td>
<td>meas</td>
</tr>
<tr>
<td>$V_T$</td>
<td>V</td>
<td>2.09</td>
</tr>
<tr>
<td>$R_{on}^{sp}$</td>
<td>Ωmm$^{-2}$</td>
<td>0.331</td>
</tr>
<tr>
<td>$BV_{DS}$</td>
<td>V</td>
<td>85.5</td>
</tr>
<tr>
<td>$h_{FE}$</td>
<td>-</td>
<td>96.4</td>
</tr>
<tr>
<td>$f_T$</td>
<td>MHz</td>
<td>-</td>
</tr>
<tr>
<td>$U_A$</td>
<td>V</td>
<td>405</td>
</tr>
<tr>
<td>$R_{sp}^{pinch}$</td>
<td>kΩ/□</td>
<td>2.4</td>
</tr>
<tr>
<td>$R_{SN}$</td>
<td>Ω/□</td>
<td>17.2</td>
</tr>
</tbody>
</table>

Order RSM is taken to model a response:

$$R = a_{00} + \sum_{i=1}^{N} a_{i0} x_i + \sum_{i=1}^{N} \sum_{j=i+1}^{N} a_{ij} x_i x_j + \sum_{i=1}^{N} a_{ii} x_i^2,$$

(2.7)

where $R$ is the response (e.g. $h_{FE}$), $a_{ij}$ a RSM model term, $x_i$ a (normalized) process parameter and $N$ the total number of process parameters. In order to obviate a long listing of all RSM coefficients, we present the results in reduced form, in figure 2.6. In this figure, the relative contribution of each linear and quadratic model term is shown; for each RSM, all contributions add up to 100%. Presented this way, the ‘signature’ plots can be easily compared in order to determine the dominant process parameters and trade-offs. The cross-terms are small and are not shown, for the sake of simplicity. The dashed lines indicate the root mean squared error between the numerical experiments and the RSM, and are a measure of the quality of the employed RSM. For instance, $V_T$ is described more accurately by its RSM than $h_{FE}$.

From the figure, the trade-off between DMOST and NPN device parameters is obvious: the signature of $V_T$ is an image of the signature of $h_{FE}$ and $f_T$. The NPN current gain and cut-off frequency cannot be considerably improved without
Figure 2.6: ‘Signature’ of the process parameters. Shown are the relative model terms of normalized RSMs for $V_T$, $R_{on}^{sp}$, $h_{FE}$ and $f_T$. The cross-terms are small and omitted.
lowering the VDMOST threshold voltage. However, the trade-off can be avoided to some extent by using the exclusive relation between $V_T$ and $T_{ox}$. The epi-layer concentration is of prime importance for $R_{on}^{sp}$, and to a lesser extent for $f_T$. Within the experimental ranges, $f_T$ appears to be mainly dependent on the emitter-base configuration, and less on the epi-layer parameters which determine the collector. This gives some room to optimize $R_{on}^{sp}$ by adjusting $N_{epi}$, insofar as it is not set by other devices. $L_G$ and $T_{ox}$ only affect the VDMOST behavior, and can be used to optimize the device independently from the NPN. The sensitivity of $R_{on}^{sp}$ for $L_G$ variations, however, is low.

Next, in the optimization phase, we only consider the process parameters which are exclusively related to the VDMOST and the NPN to ensure that other devices, such as the CMOS transistors, remain unaffected. These parameters are the SP dose $\Phi_{SP}$ and anneal time $T_{SP}$ and the drawn gate length $L_G$.

First, DEBORA calculates an optimum using the RSMs. Next, this solution is used as a starting point for an LM iteration by PROFILE. Note that due to the limited quality of the RSMs, the experiment at this starting point gives slightly different device parameter values. The VDMOST parameters were given twice as much relative weight with respect to the NPN parameters, to reflect the importance of the VDMOST device. The Early voltage received a weight twice as small as the current gain and cut-off frequency.

The results are shown in table 2.5. The optimum parameter settings, calculated by DEBORA, are given in the ‘DEBORA’ column; the results after subsequent fine-tuning with PROFILE are shown in the ‘+PROFILE’ column. In the lower part of the table, the response values are given. ‘RSM opt.’ denotes the values of the
Table 2.5: Optimization results. Of the device parameters are shown the RSM values, the actual values and the optimum found by PROFILE.

<table>
<thead>
<tr>
<th>process parameter</th>
<th>units</th>
<th>default values</th>
<th>DEBORA</th>
<th>+PROFILE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Phi_{SP}$</td>
<td>cm$^{-2}$</td>
<td>$7.5 \times 10^{13}$</td>
<td>$8.0 \times 10^{13}$</td>
<td>$9.0 \times 10^{13}$</td>
</tr>
<tr>
<td>$T_{SP}$</td>
<td>min.</td>
<td>480</td>
<td>412</td>
<td>391</td>
</tr>
<tr>
<td>$L_G$</td>
<td>$\mu$m</td>
<td>10</td>
<td>15</td>
<td>10.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>device parameter</th>
<th>units</th>
<th>target values</th>
<th>RSM opt.</th>
<th>start.</th>
<th>LM opt.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_T$</td>
<td>V</td>
<td>2.25</td>
<td>2.20</td>
<td>2.20</td>
<td>2.32</td>
</tr>
<tr>
<td>$R_{on}^{sp}$</td>
<td>$\Omega$mm$^2$</td>
<td>0.20</td>
<td>0.426</td>
<td>0.356</td>
<td>0.320</td>
</tr>
<tr>
<td>$BV_{DS}$</td>
<td>V</td>
<td>$&gt;80$</td>
<td>-</td>
<td>88.3</td>
<td>88.5</td>
</tr>
<tr>
<td>$h_{FE}$</td>
<td></td>
<td>100</td>
<td>105</td>
<td>99.4</td>
<td>95.5</td>
</tr>
<tr>
<td>$f_T$</td>
<td>MHz</td>
<td>350</td>
<td>353</td>
<td>361</td>
<td>367</td>
</tr>
<tr>
<td>$U_A$</td>
<td>V</td>
<td>450</td>
<td>473</td>
<td>409</td>
<td>369</td>
</tr>
</tbody>
</table>

Optimum in the RSMs. As this solution is used as a starting point for PROFILE, the corresponding values of the device simulations are available and shown in column ‘start.’. PROFILE slightly improved $R_{on}^{sp}$ at the expense of larger deviations of the other parameters.

As can be seen, the optimum is located near the boundary of the experimental ranges. $T_{SP}$ is reduced and $\Phi_{SP}$ is increased to create a smaller base while keeping the threshold voltage acceptable. This also results in a smaller VDMOST channel length and improved linear characteristics. However, the resulting steeper SP profile will give a greater sensitivity of the device parameters to process variations, especially of $V_T$. 
2. Device optimization techniques
Chapter 3

Determination of doping profiles

In this section, we introduce two novel nondestructive dopant profiling methods and apply them to lateral doping profiles of double diffused MOS devices. Both methods require CV measurements only.

After an introduction to existing doping profiling measurement methods, we first present a direct method which determines the lateral doping profile along the silicon/oxide interface by using a threshold voltage formulation. It is essentially a 1D method and allows a fast inspection of the lateral out-diffusion of doping profiles. Examples are given of lateral channel doping profiles of a DMOS transistor. The method is shown to be sensitive and is, therefore, particularly suited for monitoring in-line processes, or quick profile evaluation during process development.

The second method is based on the inverse modeling approach and not essentially one-dimensional. Instead of using parameterized analytical or numerical models for the doping profiles, a spline representation is used. This allows us to describe arbitrary doping profiles with a limited number of coefficients. As no simplifying assumptions are necessary, this method is more accurate than the first one, but it requires considerable computational effort.

3.1 Introduction

The measurement of two-dimensional (2D) doping profiles as part of the complete 2D characterization of device structure and geometry is a critical issue in IC technology. The rapid decrease of device dimensions into submicron levels is not accompanied by the development of fast and reliable 2D characterization methods. As the electrical characteristics of small devices suffer from 2D effects, the need
for a good description and prediction of both structure and geometry is great. In the absence of methods capable of measuring accurate 2D doping profiles in actual devices, optimization of small devices is a difficult matter, as we need to rely on insight gained by modeling. Whereas small devices would benefit from 2D characterization, the need is perhaps even more obvious for essentially two-dimensional devices such as the DMOS transistor under consideration in this thesis.

In literature, over the years, numerous destructive and nondestructive measurement techniques have been proposed and studied, both one- and two-dimensional. In the following sections, we first give a short review of existing profiling methods. The review is intended to be typical rather than complete; for a more extensive overview the reader is referred to the excellent review article of Subrahmanyan [158] and the references therein. Several criteria can be used to group published methods. In this chapter, we make a distinction between destructive and nondestructive methods, as the latter can, in general, be used for actual devices or specially designed test structures without requiring additional wafer processing or sample preparation. Therefore, nondestructive methods are attractive for in-line process characterization, especially when electrical measurements can be used. A further distinction is made between vertical, lateral and full 2D profiling methods, as they reflect the increasing difficulty and decreasing accuracy of doping profiling in general. Other classifications could result from the following considerations:

- A distinction must be made between chemical and electrical dopant profiles. The profiles can be very different, depending on the processing conditions.

- The complexity of the measurement method determines its practical applicability. A complex method cannot be routinely used as a characterization tool for in-line processes.

- Direct methods map the measurements directly into a doping or carrier profile by using a unique relation between the profile and measurement data, e.g., the concentration-dependent depletion layer width in the capacitance voltage techniques [113]. If no such relation exists, inverse methods are used in which a doping profile is iteratively improved by fitting simulated to experimental data.

Before we present the novel methods, we first give an introduction to existing doping profiling methods.
3.2 Existing doping profile measurement techniques

3.2.1 Destructive methods

Destructive methods in general proceed by removing thin layers of material or by beveling the sample, followed by an analysis of either the removed layer or the remaining sample.

Vertical

The most common technique to resolve atomic doping profiles is secondary ion mass spectrometry (SIMS). The method is based on mass spectrometry: A rastering primary beam of, e.g., oxygen sputters a crater (typically 250×250 μm) of which the sputtered ("secondary") ions are analyzed by mass spectrometry. After the measurement, the crater depth is determined and the average sputtering rate is used to reconstruct the depth-dependent doping profile from the time-dependent response, which results in some uncertainty in the depth scale (5–10%).

Depending on the calibration method, an accuracy of 10 to 30% is achieved for the absolute concentration scale. The lowest detectable concentration level with this accuracy is approximately $1 \times 10^{16}$ cm$^{-3}$. Despite the uncertainty in the absolute doping level, SIMS is very useful, because of its high relative accuracy and reproducibility. The resolution is $< 5$ nm and the relative accuracy in both the measured doping level and depth scale are $< 10\%$. If measurements are done in one batch, an even better accuracy is obtained.

Whereas SIMS measures the chemical doping profile, spreading resistance profiling (SRP) measures the electrically active part of the profile. Originally, anodic sectioning (oxidation + selective etching) was used to remove thin layers of material, after which the resistance of the remaining sample was measured with electrical probes. Alternatively, the resistance can be measured on a bevel, which requires much simpler sample preparation and gives both a higher accuracy and better reproducibility. The resolution is then determined by the bevel angle in combination with the step increment in the probe placement, which are both mechanical parameters.

Actually, SRP measurements give carrier concentrations rather than electrical active dopant profiles, which need not be the same, due to carrier diffusion effects. This results in an error when measuring steep profiles. Therefore, many authors apply the inverse modeling approach in which the Poisson equation is solved to it-
eratively calculate the active profile from the measured carrier profile or resistance data, see, e.g., Choo et al. [20]. To avoid instabilities during the solution of Poisson's equation, and thus to ensure convergence of the procedure, data smoothing must be applied. This limits the accuracy.

A resolution of 5 nm is achieved and the method has an excellent sensitivity, especially for low concentrations in contrast to, e.g., SIMS. An absolute depth scale is easily obtained by using an oxide capping layer.

**Lateral**

The literature on lateral profiling techniques is less abundant, as the use of common measurement techniques for lateral profiling is difficult. Lateral profiling methods are essentially 1D by assuming that the vertical profile is constant. In combination with vertical profiling methods, however, they may provide useful information to check and calibrate computer models.

Most published lateral profiling methods are based on electrical measurements using, e.g., Scanning Tunneling Microscopy (STM) probes to step along the surface and to collect electrical data [67,148,174]. The only method to resolve atomic profiles was published by Von Criégern et al. [31] who used SIMS in samples sectioned perpendicular to the surface and parallel to the mask edge. Their method required careful sample preparation and was only applied to high-dose arsenic doping distributions, to ensure a high enough signal intensity through the relatively high ion yield of arsenic.

**Two-dimensional**

In recent years, many publications about 2D destructive profiling methods can be found. Many of them are based on 1D techniques such as SIMS [21, 51, 98] and SRP [129,171,172]. By using clever device structures and by careful sample preparation, such as, e.g., a double bevel, the poor lateral resolution of these techniques can be circumvented.

Other methods make use of carrier-concentration-dependent etching rates to obtain sample thickness variations which are imaged by, e.g., STM [162]. Transmission electron microscopy (TEM) is also used to detect thickness variations which appear as fringes (isoconcentration contours) in the image [97, 100]. The sensitivity of these methods is limited to approximately $5 \times 10^{17}$ cm$^{-3}$ and the accuracy and reproducibility depends on the etching and calibration procedure.
3.2. Existing doping profile measurement techniques

More recently, atomic force microscopy and STM equipment have been used as very small and precisely positionable probes to collect electrical data, either on a cross-section or planarly [62, 145, 176]. In these methods, calibration of the electrical characteristics of the tip is necessary.

In general, most destructive 2D techniques are quite laborious and require dedicated test structures and/or expensive equipment. From this perspective, two-dimensional SRP (see e.g. Vandervorst [172]) seems to be the most attractive, because of the relative easy sample preparation and the experience with one-dimensional SRP.

3.2.2 Nondestructive methods

**Direct vertical**

Many nondestructive methods have been proposed to resolve vertical doping profiles. We here focus on electrical methods which measure charge carrier profiles rather than doping profiles. The main objective of nondestructive methods is to obtain profile data with relatively cheap measurement equipment while using test structures from in-line processes.

The most commonly used technique is the **capacitance voltage method** (CV method) which dates back to Schottky [140]. The small signal capacitance of a depletion region is measured for various depletion widths. By applying the “abrupt depletion approximation”, and the “one-sided junction approximation” in the case of $p-n$ structures, the doping profile can be directly calculated from the CV data. Several authors have considered the errors introduced by these approximations and proposed correction formulae to enhance the application range of the CV method [95, 101, 175, 183]. The sensitivity of CV methods is excellent, especially for low doping levels.

A dual method is the **channel conductance method**, as described by Ouwerling et al. [123], which measures the gate voltage dependent resistance of the undepleted part of a field effect transistor. This method requires the knowledge of the carrier mobility.

As in the MOS-CV method, the **dc-MOSFET method** scans the channel region by means of an extending depletion layer. Shannon and Buehler showed that the
width of this layer can be expressed in terms of \textit{dc} voltage changes when a constant drain current is maintained by a simultaneous variation of the gate and back gate voltage [15, 146]. The method thus requires \textit{dc} measurements only.

In the \textbf{threshold voltage method}, the back bias dependence of the threshold voltage is used to determine the channel profile, see, e.g., references 45 and 178. As does the \textit{dc}-MOSFET method, this only requires \textit{dc} measurements. To obviate a large set of measurements, Feldbaumer and Schröder used a "constant drain current" threshold voltage, which, they argued, differs only by a constant offset of the real threshold voltage, and used this in an automated setup to quickly extract the profile [45].

\textbf{Direct lateral}

Only a few nondestructive direct methods to measure the doping profile along the surface rather than into the depth have been published.

The method described by Lee and Dutton [91] is based on a gate voltage dependent inversion layer width. When the gate is positioned above a laterally varying doping profile, part of it will invert, depending on the gate voltage. Lee and Dutton used a threshold voltage formula to calculate the inverted doping level at a specific bias. They determined a corresponding coordinate by measuring the channel conductance from which they calculated the inversion channel width. This width is then interpreted as a lateral coordinate. For this calculation, the surface mobility must be known.

Rössner \textit{et al.} and Uchida \textit{et al.} also used a threshold voltage formulation, but in combination with overlap capacitance measurements to determine a lateral coordinate [133, 168]. They biased a MOSFET in accumulation, thereby inverting the tail region of the source and drain doping profiles. Then, the tail region is electrically connected to the back gate region and it is shielded from a gate-drain and gate-source overlap capacitance measurement. At increasing gate voltages, the drain and source profiles gradually invert and an even larger part is shielded. Rössner \textit{et al.} and Uchida \textit{et al.} used this bias dependence of the overlap capacitance to calculate the length of the inverted part of the profile, and interpreted this as the lateral coordinate for the extracted profile. They applied their method to sub-micron source and drain profiles, which appears to be awkward, due to the high 2D nature of the device geometry.
3.2. Existing doping profile measurement techniques

3.2.3 Inverse methods

Several authors have proposed inverse methods to overcome the limitations of direct methods. In such methods, the doping profile is determined by back-calculating the experimental data (e.g. CV data).

The use of nonlinear parameter estimation techniques to solve inverse problems in semiconductor device physics was introduced as Inverse Modeling (IM) by Ouwerling et al. [124]. A general introduction to IM is given by Crans [30]. A device simulator is used to calculate the electrical characteristics of a device with a parameterized doping profile. With an optimizer, the profile parameters are then iteratively adjusted so as to minimize the difference between simulation and measurement.

IM is more general than direct methods. For instance, it allows the simultaneous extraction of other device parameters, such as oxide thickness or surface charge. Jimenez-Tejada, for instance, extracted platinum-induced deep levels in nonabrupt $p$–$n$ junctions simultaneously with the impurity profiles [73]. Also, simplifying approximations, such as the abrupt depletion layer or the absence of carrier spilling effects, are not necessary. Toyabe claimed sub-Debye length resolution in the surface region of MOS structures, although convergence was awkward [167]. Osse showed with a synthetic example that the determination of ultrashallow MOS doping profiles is feasible [119]. Also, extraction of profiles in more than one spatial dimension is possible, as first demonstrated by Ouwerling, who extracted a Junction CCD profile [120, 121].

A potential source of error in IM is the choice of the parameterized form of the doping profile. Whereas analytical functions, e.g., a Gaussian, would be appropriate for deeply diffused low concentration profiles, they are hardly so for state of the art IC technologies. Numerical process simulation tools give more reasonable results, but are still only able to model a limited set of profiles. To overcome these limitations, Ouwerling proposed to iteratively solve a dedicated linear least-squares systems which describes the internal electrical state of the device. Externally measured changes in the terminal capacitance, for small increments of the $dc$ terminal biases, can be expressed by internal changes of carrier concentrations. The carrier concentrations can be calculated by solving Poisson's equation, using an initial guess for the doping profile. By discretizing the doping profile and measuring $CV$ data for a large set of $dc$ biases, a linear least-squares system can be constructed. To improve the initial guess of the doping profile, the simulated
capacitances are replaced by the measured ones, and the system is solved to calculate a profile update. This procedure is repeated until simulated and measured data match. This scheme has been shown to work, although it is difficult to implement from a practical point of view. To arrive at a least-squares system, it must be assumed that the capacitance changes are small. This requires either a large set of measurements, and hence a large computational effort, or a careful measurement setup minimizing the number of bias points without losing profile information.

To generalize IM to arbitrary profiles, Khalil proposed to use spline functions for the doping profiles [80, 81] thus yielding a large, but still manageable, set of parameters. Spline functions, however, are defined by using ‘knots’, their own ‘mesh’, which require careful placement. Although these positions could, in principle, be included in the extraction procedure, this results in at least twice as many parameters to determine. Therefore, Khalil proposed to guide the knot placement as much as possible by information of the device already available, such as SIMS profiles and effective channel lengths. In a subsequent paper, Khalil presented a heuristic method to determine knot positions for 1D MOS profiles [82].

To summarize, the main advantages of inverse methods are:

- No simplifying assumptions are required to express the doping profile in the measured variable;
- They are not restricted to one spatial dimension.

The main disadvantages are:

- The sensitivity for high doses is limited;
- A relatively large computational effort and the use of sophisticated numerical software tools are required.

3.3 Threshold voltage profiling method ("$V_T$ method")

In this chapter, we present a fast nondestructive direct lateral doping profiling method that uses CV measurements and a threshold voltage formulation. The method uses a wide gate test device and standard CV measurement equipment, and is therefore very useful to quickly examine the lateral out-diffusion of moderate doping profiles, such as the back gate channel profile of the DMOS transistor.
After an introduction of the method, we first give a numerical example to compare the extracted profile to the actual one. Next we discuss the limits and the validity of the method before we give measurements of actual doping profiles in a lateral DMOS transistor.

3.3.1 Principle of the method

It is known that gate overlap capacitances in a MOSFET are bias dependent, especially for a lightly doped drain structure, see, e.g., De Graaff and Klaassen [53]. For instance, when a MOSFET is in saturation, the drain gate voltage induces a depletion layer in the drain under the gate, thereby decreasing the gate drain capacitance.

A similar situation occurs when the MOS is biased in the accumulation regime, as illustrated in figure 3.1. This schematic representation shows the depletion layers of a pMOSFET in accumulation for two different gate biases. The overlap capacitances $C_{GD}$ and $C_{GS}$ obviously depend on the depletion layer widths and are therefore bias dependent. Moreover, at increasing gate voltages, the lower part of the lateral drain and source profile inverts, and will be electrically connected to the grounded bulk. This inversion layer shields the depletion layer from the gate, which even further decreases the overlap capacitance. Obviously, the bias depend-
ence of the overlap capacitance depends on the lateral doping profile.

The $C_{GS}(V_G)$ and $C_{GD}(V_G)$ data can be converted to a lateral doping profile as follows. Consider the source and drain region as a series of MOS capacitors (see enlargement in figure 3.1), each with constant but different "bulk" doping levels. At a given gate voltage, part of these MOS capacitors will be inverted with the inversion layer connected to ground, and will thus not contribute to the measured overlap capacitance. The overlap capacitance therefore contains information about the penetration depth of this inversion layer. Next, the threshold voltage formula can be used to calculate the maximum doping level that is actually inverted. In the $V_T$ formula we therefore replace $V_T$ by $V_G$:

$$V_G = V_{FB} + 2\phi_F + \gamma \sqrt{2\phi_F} \text{ with } \gamma = \sqrt{2\epsilon_{si} q N_A / C_{ox}}$$ \hspace{1cm} (3.1)

in which $V_{FB}$ is the flat band voltage, $\phi_F$ the bulk potential $kT/q \ln(N_A/n_i)$, with $n_i$ the intrinsic carrier concentration; $\gamma$ is the body effect coefficient, $N_A$ the doping level and $C_{ox}$ the oxide capacitance per square meter ($= \epsilon_{ox}/t_{ox}$).

The overlap capacitance depends on the depletion layer widths of all non-inverted MOS capacitors, which in turn depends on the bias condition. At each bias, they have to be recalculated in order to determine the required penetration depth of the inversion layer.

To avoid the calculation of depletion widths, we use the gate bulk capacitance $C_{GB}$ rather than $C_{GS}(V_G)$ or $C_{GD}(V_G)$, see figure 3.1, assuming that the source and drain profiles are equal. Now, a simple scheme results to resolve the lateral profile. The equations for, e.g., a $p$-type profile are then given by:

$$N_A^i = \frac{C_{ox}^2}{2\epsilon_{si} q} \frac{(V_G^i - V_{FB} - 2\phi_F)^2}{2\phi_F}$$ \hspace{1cm} (3.2)

and

$$L^i = L^{i-1} + \frac{C_{GB}^{meas,i} - C_{GB}^{meas,i-1}}{2C_{ox} W},$$ \hspace{1cm} (3.3)

with $N_A$ the doping level to extract, $L$ the relative lateral axis, $W$ the width of the MOS capacitor, and $i$ is the index of the $i$'th $C_{GB}(V_G)$ measurement.

To summarize, we use $C_{GB}(V_G)$ measurements to determine the relative lateral position where the strong inversion condition in the source and drain is just met, and we use the threshold formula to calculate the doping level at that position.
3.3.2 Synthetic example

To guide the discussion about the limits and validity of the assumptions underlying the lateral $V_T$ profiling method in the next section, we first analyze a synthetic device. The advantage of a numerical study is that the exact structure and actual carrier concentrations at each bias are known.

The structure we study is essentially a gated diode, and is shown in figure 3.2. The oxide thickness (70 nm) and maximum doping level ($1 \times 10^{18}$ cm$^{-3}$) are chosen to resemble the back gate doping levels of the VDMOST shown in figure 1.2. For comparison, three different lateral diffusion lengths are used: 0.3 $\mu$m, 0.75 $\mu$m and 2 $\mu$m.

![Diagram of gated diode](image)

Figure 3.2: Schematic illustration of the gated diode which is used to calculate synthetic CV data, in order to evaluate the lateral profiling method. Three different out-diffusion lengths $L_{OV}$ were used: 0.3, 0.75 and 2 $\mu$m.

The simulation program MEDICI [164] is used to calculate the gate-bulk ($C_{GB}$) capacitance data. Figure 3.3 gives the simulated $C_{GB}(V_G)$ characteristics for the 2 $\mu$m out-diffused profile. At higher $V_G$, a larger part of the profile will be inverted, which is reflected by an increasing $C_{GB}$. When the entire drain profile has been inverted $C_{GB}$ obviously saturates.

Using equations 3.2 and 3.3 and the synthetic $CV$ data, we can reconstruct the lateral doping profile as shown in figure 3.4. To pin the extracted axis, we use the gate voltage at which the entire profile has just been inverted, i.e. $V_G \approx 15$ V. The extracted profiles match the original ones very well, except near the gate edge and the lateral junction.
3. Determination of doping profiles

Figure 3.3: Simulated $C_{GB}(V_G)$ characteristics of the gated diode with a 2 μm out-diffusion length.

3.3.3 Limits and restrictions

The concept of the lateral $V_T$ method is essentially 1D, a condition which is not always met. Rösner et al. [133] and Uchida et al. [168] applied their implementation to submicron structures for which the 1D assumption is difficult to maintain. For larger structures, however, such as the DMOS channel profiles in this thesis, the method gives accurate results and can be used to quickly assess the amount of lateral out-diffusion using standard $CV$ measurement equipment.

In this section, we discuss the limits and the validity of the assumptions of the method. These are:

- Two-dimensional effects near the junction and the gate edge,
- The relative extracted axis,
- The sensitivity and resolution.

Near the junction and gate edge

As the method is essentially 1D, it obviously breaks down near the lateral junction and the gate edge.

Near the lateral junction, charge is shared by both junction and gate and, moreover,
Threshold voltage doping profiling...

Figure 3.4: 'Original' and 'measured' lateral p-type doping profiles for three different out-diffusion lengths: 0.3, 0.75 and 2 µm, here aligned on the lateral junction to allow easy comparison. The original profiles are represented by symbols, the extracted profiles by lines. On top of the figure, the position of the gate edge with respect to the profile is illustrated.

the net doping varies rapidly over the Debye length thus impeding the interpretation of the carrier concentration as a good measure of the activated dopant density [113]. This is inherent to all direct CV methods which do not use correction formulae. In effect, the \( V_T \) method overestimates the doping near the junction and is not able to clearly extract the junction.

Near the gate edge, the electric field lines diverge and use of the 1D threshold voltage formula is no longer justified. As a result, we observe a pronounced — nonphysical — pile-up of the extracted profile which is actually helpful to pin the measured profile to the gate edge. In order to discriminate between this 2D effect and actual doping effects, it is worth studying this feature in more detail. To study the effect of the gate edge on the extracted doping profile, we give a simplified analysis based on the charge neutrality assumption.

The threshold voltage roll-off of small geometry MOSFETs has been described
3. Determination of doping profiles

by considering that the charge on the bottom of the gate is compensated by a —
nonuniformly distributed — bulk charge. In figure 3.5, we follow the same
procedure. Charge neutrality requires that the charge on the gate is compensated by
charge in the silicon. The total capacitance can be thought of a pure parallel plate
capacitor $C_{ox}$ in parallel to a fringing capacitance. This fringing capacitance can
be regarded as the sum of three components [79]: the side wall capacitance $C_{sw}$,
the lower edge capacitance $C_{le}$ and the top capacitance $C_{top}$. $C_{le}$ affects the charge
distribution near the gate edge. So, to arrive at a more accurate expression for $V_T$,
we need to consider the charge on $C_{le}$. We now make the following assumptions:

- Charge balance is assumed on both sides of the point of strong inversion, so
  we can restrict the analysis to the noninverted part of the structure.

- The charge on the lower edge capacitance $C_{le}$ is assumed evenly distributed
  over the bottom and the side wall of the gate. To obviate a calculation in-
  cluding side wall effects, we roughly assume that half of the charge must be
  considered in the analysis.

- The voltage drop across the oxide is considered to be constant, which is a
  reasonable assumption for relatively large gate voltages (the silicon can only
  account for $\approx 1$ V before inversion sets in). This allows us to easily calculate
  the charges on $C_{ox}$ and $C_{le}$.

Writing down the threshold voltage in terms of charges we obtain:

$$V_T = V_{FB} + 2\phi_f + \frac{Q_B}{L'C_{ox}},$$

(3.4)

in which $Q_B$ is the total charge per meter in the bulk and $L'$ the length of the non-
inverted part of the gate, both are illustrated in figure 3.5. We approximate $Q_B$
by:

$$Q_B = Q_{BD}^1 L' + \frac{V_{ox} C_{le}}{2},$$

(3.5)

where $Q_{BD}^1$ is the gate oxide capacitance charge per square meter. Using 3.5 in 3.4
we obtain:

$$V_T = V_{FB} + 2\phi_f + \frac{Q_{BD}^1}{C_{ox}} + \frac{V_{ox} C_{le}}{2L'C_{ox}}$$

(3.6)

$$= V_{FB} + 2\phi_f + V_{ox} \left( 1 + \frac{C_{le}}{2L'C_{ox}} \right)$$

(3.7)
$C_{le}$, in general, depends on the shape of the gate edge but can be calculated analytically for a gate with a rectangular edge [79]. It appears to be almost independent of gate and oxide thickness:

$$C_{le} \approx 0.613 \frac{\varepsilon}{\pi},$$  \hspace{1cm} (3.8)

for silicon $C_{le} = 6.7$ pF/m. In equation 3.6 we observe that, near the gate edge, $V_T$ is higher than the 1D equivalent and we therefore overestimate the doping level.

The extent of the influence of $C_{ox}$ under the gate may be examined by calculating the distance from the gate edge where the doping calculated from equation 3.6 differs by 10% from the one calculated from equation 3.1. If we assume a large $V_T$, so we may neglect $V_{FB}$ and $\phi_f$, we find after some calculations:

$$L' \approx 10 \frac{C_{le}}{C_{ox}},$$  \hspace{1cm} (3.9)

which can be rewritten, by using 3.8, as:

$$L' \approx 2 \cdot t_{ox}.$$  \hspace{1cm} (3.10)
In our example, $t_{ox}$ is 70 nm, and we therefore find $\approx 140$ nm which is actually somewhat larger than the pile up of the extracted profiles shown in figure 3.4. Note that in this simple model, it is the oxide thickness that determines the extent of the influence of $C_{le}$ and not the lateral doping profile. Therefore, the effect of gate edge on the profile extraction is relatively large for smaller out-diffusion lengths, as evident in 3.4. As a 'rule of thumb', we find that the lateral diffusion must exceed the oxide thickness considerably to prevent fringing effects from disturbing the extraction the measured profile.

Although this analysis becomes inaccurate when the inversion layer approaches the gate edge very closely (for instance, equation 3.6 then predicts $V_T \rightarrow \infty$, which is obviously not the case), this allows us to qualitatively understand why the extracted profiles pile up before the whole gate is inverted.

Once the entire profile beneath the gate is inverted, the measured capacitance $C_{GB}$ will not increase further for even higher gate voltages. When using measurement data from this regime, the extracted profile will pile up even stronger.

**Relative axis**

As is the case with profiling the base of bipolar transistors by means of $CV$ measurements, we find a relative doping axis. However, the pile-up allows us to pin the extracted profile rather accurately to the gate edge.

**Detection limits and resolution**

The lower and higher limit on the extracted doping values are determined, roughly, by the background doping level and the electrical breakdown of the gate oxide, respectively. The maximum doping level that can be inverted before oxide breakdown occurs is $\approx 3 \times 10^{19}$ cm$^{-3}$. Care must also be taken to avoid bias conditions, such as deep depletion, at which hot carriers are injected in the oxide (see e.g. Sze [159]), as these affect the $CV$ measurement results.

The resolution of the doping axis is determined by the number of $C_{GB}(V_G)$ measurements, as each measurement yields a coordinate. Obviously, $C_{GB}$ must be large enough to enable accurate measurement of small increments. Large $C_{GB}$ values, however, are easily obtained by designing a device with a large gate width. This relieves the demand for such advanced measurement techniques as were, e.g., necessary in Williams [174], who used the probe of an atomic force microscope to collect $CV$ data. For instance, if the $C_{GB}(V_G)$ characteristics of figure 3.3 were measured at 100 mV intervals, the total number of data points in the extracted pro-
file would be over 100. The doping profile had a lateral diffusion length of 2 \( \mu \text{m} \), so the resolution would then be approximately 20 nm.

**Summary**

To summarize, we note the following advantages of the \( V_T \) method:

- **Ease of use:** no additional processing or sample preparation is required;
- **Using a direct formulation and CV measurements only,** the method is quick;
- **Standard CV measurement equipment can be used;**
- **As with all CV methods,** the sensitivity is excellent.

The limits and restrictions are:

- **Dopant range:** typically \( 1 \times 10^{15} - 1 \times 10^{19} \ \text{cm}^{-3} \);
- **Resolution of the dopant axis:** is determined by the gate width and the minimal increments in \( V_G \);
- **The lateral diffusion length** must exceed the oxide thickness considerably to avoid 2D effects;
- **The \( x \) axis is a relative axis,** but can be fixed by aligning the sharp pile-up in the measured profile to the gate edge.

Using equation 3.6 in a parameterized form, the method can be applied to small out-diffusion lengths [69]. This, however, requires the knowledge of additional parameters which depend on the actual fringing capacitance and thus are, in general, technology dependent.

### 3.3.4 DMOS channel profile measurements

**Test structures**

Two test structures have been designed and processed in order to measure the lateral channel profile in a DMOS transistor. The devices were included in two test batches (EC1389 and EC1558) of the L303 process currently in-line at Philips. The doping profiles which are used to define the devices are named:
3. Determination of doping profiles

- SN, 'shallow n': source and drain (also used as emitter in the NPN)

- SP, 'shallow p': back gate (also used as base in the NPN)

- SP2, second 'shallow p': under back gate contact (also used under base contact in NPN)

The first device is a lateral DMOS transistor which is used, rather than a vertical one, to avoid 'averaging' of two profiles: due to shadowing effects during implantation, the two net channel profiles in a VDMOST may differ. A cross-section of part of the device is shown in figure 3.6. For increasing gate voltages, the channel region will invert from the drain side and the channel profile is extracted by measuring the gate drain capacitance $C_{GD}$. The method fails to resolve the channel profile near its top. Then, the entire channel region is inverted and the drain is connected to the grounded source which impedes reliable $C_{GD}$ measurements. Also, we have no information about the absolute coordinate, as there will be no characteristic pile-up in the measurements in the absence of a gate edge above the channel profile.

Therefore, a second structure is included in the second batch (ECI558) which differs from the first one only by the exclusion of the source. It is, in essence, a gated diode, such as the synthetic examples given in section 3.3.2. The $C_{GD}$ measurement is now not hindered by a short to ground, and the p-type back gate profile can be "probed" straight to the gate edge. Due to a gate-reoxidation step, however, the oxide thickness near the gate edge is thicker, which limits the measurement accuracy within approximately 100 nm of the gate edge. Ideally, the measured profiles from both structures should match in the low dose part, provided that the source implant did not affect the net channel profile.

Both batches contain nominal devices, but in batch ECI389, two additional devices were included by varying the nominal SP dose of $7.5 \times 10^{13}$ cm$^{-2}$ by 10%. Both test structures have a drawn gate length of 6 $\mu$m. A gate width of 1.6 cm was used, split in 16 parallel fingers of 1 mm length each. The large channel width was chosen by assuming, initially, a channel length of 1.5 $\mu$m so that a $C_{GD}$ variation of about 8 pF was expected while scanning the complete channel. This is large enough to measure small capacitance increments, and thus to measure the lateral channel profile with high resolution. The gate voltage was ramped at 50 mV steps. With the device threshold voltage of 2.5 V, this results in approximately 20 data points along the channel doping profile.
Figure 3.6: Detail of the LDMOST test device which was included in the test batches ECI389 and ECI558 of the L303 process currently in-line at Philips. In batch ECI558, a gated diode was included which only differs from the LDMOST by the exclusion of the source, was also included. By measuring $C_{GD}(V_G)$, the channel region can be probed.

Measurements

The $CV$ measurements were carried out using an HP4274A instrument. Figures 3.7a and 3.7b show the $C_{GD}(V_G)$ measurements for both structures. The structure without source shows qualitatively the same $CV$ curve as in the synthetic example given in section 3.3.2. At $V_G \approx 8$ V, $C_{GD}$ saturates, indicating that the entire profile beneath the gate is inverted.

However, the measurement of the structure with a source fails near the device threshold voltage of approximately 2.3 V, due to the short of the drain to the grounded source. In this measurement we identify three regions of interest, indicated in figure 3.7a:

1. Depletion of the drain;
Figure 3.7: $C_{GD}(V_G)$ measurements which were used to calculate the lateral doping profiles in figure 3.8. a) LDMOST structures with dose variation in batch ECI389. b) LDMOST and adjacent gated diode from batch ECI558.

2. Partly inverted channel;

3. Totally inverted channel.

Obviously, the $V_T$ method can only be applied in the second region. In the first region, the drain is in depletion. $C_{GB}$ cannot be considered a pure oxide capacitance, and can therefore not be interpreted as a coordinate axis. In the third region, the complete channel is inverted and the $CV$ bridge is no longer capable of measuring
$C_{GD}$, as the signal is grounded to the source.

In figure 3.8a, the extracted doping profiles are shown. From batch ECI389, three channel profiles were measured and, as can be observed, the sensitivity to the dose variation is excellent. Figure 3.8b shows the profiles from batch ECI558.

Figure 3.8: a) Comparison of the extracted profiles of the LDMOST channel for process variations. b) Comparison of the extracted profiles of the LDMOST and the gated diode ("LDMOST without source"). As the channel profile is in principle the tail of the gated diode profile and its absolute coordinate is not known, we shifted it an attempt to overlay both profiles. Obviously we did not succeed, which indicates that the source implantation affected the net channel doping level in the LDMOST.

If the source implantation did not affect the net channel doping level, the profiles
from both test structures must coincide. Obviously, both profiles cannot be overlaid, indicating that the source implantation did affect the net dopant distribution in the channel region. It cannot be explained by compensation effects due to the out-diffused phosphorus, or by a boron redistribution, as this would result in a less shallow channel profile instead of the measured shallower one.

However, during processing, it was concluded from other test structures, that in batch ECI558, a tail of the phosphorus implant reached through the polysilicon and gate oxide into the bulk, thereby compensating the boron. The net channel dope in the LDMOST is then lowered by a constant amount, which gives a shallower net lateral profile. In the gated diode, however, the back gate region was masked for the SN implant, which gives the uncompensated and less shallow profile.

### 3.4 Spline profiling method

The $V_T$ method discussed in the previous section is a direct method to calculate the lateral doping profile from $CV$ measurements by using a $V_T$ formula. To avoid having to make this assumption, we can use the inverse modeling (IM) approach in which an initially assumed profile, preferably parameterized, is iteratively improved by an optimizer which matches the simulated and measured $CV$ data. Analytical or numerical models can be used to describe the doping profiles. However, a major drawback is the “stiffness” of these models. Obviously, unmodeled features of the doping profiles cannot be resolved and care must be taken to prevent these from affecting the extraction of the correctly modeled parts.

As it is still not possible to model doping profiles accurately in many practical situations, there is need for a more flexible doping model. In theory, the doping level on every grid point can be used as a parameter in the nonlinear parameter extraction. However, this would imply a large set of parameters, and it would result in an unmanageable reconstruction procedure that would require excessive computation times.

Khalil therefore recently proposed [80, 81] and used a spline representation of the doping profile to relax the number of parameters while maintaining global flexibility. In this section, we pursue this method and apply it to the extraction of a lateral doping profile of the gated diode test structure discussed in the previous section. The implementation of the method communicates with the simulators through the ‘Technology Interface Format’ (TIF) of Technology Modeling Associates (TMA) which allows us to use arbitrary TMA software. For instance, a process simulation can be used to calculate the device geometry and the initial doping pro-
file that serves as a starting guess for the spline representation.

In the first subsection we introduce the $B$-spline and spline functions. Next, we describe how we can use spline functions to set up a parameterized doping profile. Subsequently, we discuss the implementation before we show, in the final subsection, experimental results for the DMOST structures.

### 3.4.1 $B$-splines and spline functions

Spline functions are widely used in numerical analysis for, e.g., data smoothing or interpolation. They are defined by a linear combination of splines, which only take value in a restricted part of the considered domain. For this purpose, the ‘$B$-spline’ is very popular, as it can be calculated in a computational stable and attractive way. A useful and practical introduction to splines is given by Dierckx [32], whereas the mathematical foundation of $B$-splines, as well as efficient computer algorithms to calculate them, can be found in De Boor’s book [9]. Here, we only give the definition of the $B$-spline and the most important properties.

Using the same notations as Dierckx, we first define the ‘truncated power function’

$$
(x - c)^k_+ = \begin{cases} (x - c)^k_+ & \text{if } x \geq c, \\ 0 & \text{if } x < c. \end{cases}
$$

(3.11)

The $B$-spline $N_{i,k+1}$ of degree $k$ with ‘knots’ $\lambda_i, \ldots, \lambda_i + k + 1$ is defined as:

$$
N_{i,k+1}(x) = (\lambda_{i+k+1} - \lambda_i) \sum_{j=0}^{k+1} \frac{(\lambda_{i+j} - x)^k_+}{\prod_{l=0}^{k+1} (\lambda_{i+l} - \lambda_{i+l})},
$$

(3.12)

which has, amongst others, the following useful properties:

- **Positivity:**

$$
N_{i,k+1}(x) \geq 0 \text{ for all } x.
$$

(3.13)

- **Local support:**

$$
N_{i,k+1}(x) = 0 \text{ if } x \not\in [\lambda_i, \lambda_{i+k+1}]
$$

(3.14)
3. Determination of doping profiles

- **Recursion:**

\[ N_{i,k+1}(x) = \frac{x - \lambda_i}{\lambda_{i+k} - \lambda_i} N_{i,k}(x) + \frac{\lambda_{i+k+1} - x}{\lambda_{i+k+1} - \lambda_{i+1}} N_{i+1,k}(x) , \]

(3.15)

*B*-splines of arbitrary degree can thus recursively be expressed in terms of the *B*-spline with the lowest degree \((k=1)\), which is simply (see equation 3.12):

\[ N_{i,1} = \begin{cases} 1, & \text{if } x \in [\lambda_i, \lambda_{i+1}) , \\ 0, & \text{if } x \not\in [\lambda_i, \lambda_{i+1}) . \end{cases} \]

(3.16)

The last property makes *B*-splines especially useful for implementation into computer codes. Figure 3.9 is an illustration of a *B*-spline of the second degree \((k = 2)\) which spans 3 knots.

![Graph of B-splines](image)

**Figure 3.9:** Illustration of a linear \(N_{2,1}\) and two quadratic *B*-splines \((N_{2,2} \text{ and } N_{0,2})\) of which one is defined using ‘coincident boundary’ knots to ‘force’ the *B*-spline to take value only inside the shown interval. The spline functions we use are a linear combination of such *B*-splines.

Given a sequence of knots, a *spline function* \(s(x)\) can be defined as a linear combination of *B*-splines:

\[ s(x) = \sum_i c_i N_{i,k+1}(x) , \]

(3.17)
in which \( c_i \) are the \( B \)-spline coefficients. The basis functions \( N_{i,k+1} \) add up to unity for all \( x \). Property 3.14 implies that only a few coefficients determine the spline function at point \( x \), depending on degree \( k \) which thus determines the local stiffness of the otherwise arbitrary tunable spline function.

To define a spline function over an interval \([a, b]\) spanned by \( g + 2 \) knots (\( g \) being the number of interior knots), boundary conditions can be imposed by using \( k \) extra coincident boundary knots at the interval ends:

\[
\begin{align*}
\lambda_{-k} &= \lambda_{-k+1} & \cdots & \lambda_{-1} &= \lambda_0 = a , \\
b &= \lambda_{g+1} = \lambda_{g+2} & \cdots & \lambda_{g+k} &= \lambda_{g+k+1} .
\end{align*}
\]  

(3.18)  

(3.19)

It can be shown that, at the interval ends, all \( B \)-splines vanish except the two outermost, which yield a simple means of applying boundary conditions:

\[
s(a) = c_{-k} , \quad s(b) = c_g .
\]  

(3.20)

A zero derivative at the interval ends is obtained by setting \( c_{-k} = c_{-k+1} \) and \( c_g = c_{g-1} \).

In summary, a spline function as a combination of \( B \)-splines is described by a sequence of \( g+2 \) knots (not necessarily equidistant) and \( g+k+1 \) spline coefficients. As the individual and easily calculated \( B \)-splines span just a few knots, only a few coefficients determine the spline function value at a certain point. This 'local stiffness' — global flexibility' feature of splines makes them attractive for describing arbitrary doping profiles with a limited number of coefficients, thus making IM feasible.

Using "tensor product splines", the attractive 1D \( B \)-splines algorithms can be used in a straightforward way to calculate 2D ,"bivariate", splines. As their implementation is simple from an algorithmic point of view, we do not give the more complex equations. We refer to De Boor [9] for an in-depth treatment and Dierckx [32] for a more practical treatment of tensor product splines.

### 3.4.2 Parameterized doping profiles using spline functions

Spline functions can be used to approximate the — logarithm — of doping profiles, and to do so we must determine the following spline parameters:

- the degree \( k \) of the \( B \)-splines,
3. Determination of doping profiles

- the knot sequence: the number and position of the knots,
- the spline coefficients.

In IM loops, we prefer a small number of parameters, and we therefore require a small number of knots and a —preferably— low degree B-spline.

A spline function constructed from higher degree B-splines require more coefficients \((g + k + 1)\). Therefore, in general, quadratic \((k=2)\) and cubic \((k=3)\) splines are used as a good compromise between the number of knots and the accuracy of the spline function.

The number and position of the knots is not a trivial problem and should be guided as much as possible by available data of the doping profiles such as SIMS profiles, effective channel lengths, etc. Smooth profiles, for instance, only need a few knots for a complete description; sharp kinks require more knots and precise knot placement near the kinks. Algorithms exist which include knot-placement and knot-insertion routines. However, the inclusion of knot insertion and placement in the extraction procedure would at least double the number of parameters, which results in unacceptable computation times for our purpose.

In order to demonstrate the flexibility of splines, we show spline representations of SIMS measurements of doping profiles in figure 3.10. Figure 3.10a is a diffused boron profile which can be accurately described with only four knots and five coefficients by using quadratic B-splines. The first interior knot is positioned near the surface, to allow a good description of the dip in the concentration. The rather complex phosphorus profile in figure 3.10b requires one more knot to approximate the kink, which gives a total of five knots and six coefficients.

In this example, the knot placement was relatively simple as we know the doping profile. During profile extraction, this is less obvious, and it is therefore stressed again that the knot placement must be guided by information already available about the profiles. Nevertheless, this example illustrates that spline functions are potentially capable of describing complex profiles with a limited number of coefficients.

3.4.3 Implementation

The spline doping profile serves as input for a device simulator. In addition to the doping profiles, other parameters, such as the gate oxide thickness and the location of the contacts, are required to complete the device description. These can either be supplied by hand or by a process simulator. Khalil used splines to calculate a
Figure 3.10: To show that spline functions can describe actual doping profiles with a limited number of coefficients, we show fits on SIMS measurements. Using quadratic B-splines as base functions, we used four noncoincident knots (five B-spline coefficients) for boron and five knots (six coefficients) for phosphorus. The knot positions are indicated in the figure.

binary doping file as input for MINIMOS [143] in which other structure parameters must be set by hand [80]. It would, however, be more convenient to use and alter process simulation structure files, in order to retain an arbitrary device geometry
and doping profiles outside the region of interest.

Our implementation of the spline doping profiling method therefore is able to filter TMA TIF files, thereby replacing the doping values by the spline function values within predefined ‘windows’ (= 2D knot sequence).

To allow easy interfacing to the TIF files, it is convenient to use two spline functions for $n$-type and $p$-type dopants, each with their own knot window, which add up to a net profile. The knot windows need not be fully overlapping, as the profiles can be spatially separated.

As $CV$ measurements only detect net carrier concentrations, care must be taken to avoid a high degree of coupling between the $n$-type and $p$-type spline coefficients determining the profiles near the junction. Depending on the order of $B$-spline used, the influence of one coefficient fortunately extends over several knots, thus preventing too high a coupling.

We also need a robust optimizer: besides possible coupling between the spline coefficients, they might also turn out to be insensitive during the nonlinear extraction. As tensor product splines are defined over a rectangular region, areas may be covered which are not probed by the $CV$ measurements, and, as a consequence, some spline coefficients may be insensitive. The optimization strategy and tools presented in chapter 2 can be applied to avoid ineffective extractions.

To summarize, in an existing device structure, the doping is replaced by two spline functions for $n$-type and $p$-type dopants respectively in predefined knot windows. The nonlinear parameter extraction adjusts the doping profile only in regions which are probed by the $CV$ measurements. In particular, our implementation, which alters TMA TIF files, has the following advantages:

- arbitrary device geometries can be used;
- simulated doping profiles outside the knot window are unaffected and do not need to be converted to a spline function.

### 3.4.4 Reconstruction of a lateral profile

In this section, we extract the lateral doping profile of the gated diode test structure, described in section 3.3.4, by matching the measured and calculated $C_{GD}$ and $C_{GB}$. The structure does not allow collection of 2D $CV$ data and we can, therefore, only extract a lateral profile. To do so, we proceed as follows:
Figure 3.11: a) Schematic drawing of the gated diode. Self-aligned to the gate are implanted: shallow p (SP; left side) and shallow n (SN; right side). The knot windows in which the doping profile is extracted are spanned by the vertical and lateral knot sequences given in Table 3.1. b) MEDICI plot of the simulated structure. A fairly complete structure is simulated to include side wall effects in the simulations.

- With TSUPREM4 we create a complete device structure, including contacts and additional layers on top of the device, in order to include side wall effects in the simulation.

- The gate oxide thickness is fixed to the value measured on a dedicated structure (69 nm) and the constant oxide trapped interface charge density is taken to be $1 \times 10^{10}$ cm$^{-2}$ [159].

- The knots and coefficients in the vertical direction are determined by 1D simulations using the calibrated boron diffusion coefficient given in Section 2.6.

- This vertical profile is then extended laterally by scaling it, thus yielding spline coefficients in the lateral direction. Note that by assuming similarity, the spline coefficients in the vertical direction scale accordingly. The construction of a 2D profile is much like the use of a complementary error function, as in section 2.6, except that we now do not have a predetermined shape of the lateral profile.
As the boron and phosphorus profiles are spatially separated, their knot windows do not overlap. We use quadratic splines and four knots laterally for both the boron and the phosphorus. The vertical and lateral knot sequences are given in table 3.1.

- During processing, it turned out that a tail of the phosphorus implant reached through the polysilicon and gate oxide into the epi-layer. This affects the $CV$ measurements and we need to account for this effect. In the gated diode, half of the gate is masked for the SN implantation. In principle, we could add extra knots near the middle of the gate to account for this effect. But as the tail is parasitic, and we are primarily interested in the source/drain and back gate profiles, we content ourselves with describing this effect by adding a simulated phosphorus tail. It is appropriately scaled and added to the background epi-layer doping level of $4 \times 10^{15}$ cm$^{-2}$. We need two additional parameters, i.e., a scaling factor and a lateral characteristic length.

- The overlap capacitances $C_{GB}$ and $C_{GD}$ are calculated with MEDICI, using the AC-analysis option, for $-35 \text{ V} < V_G < 35 \text{ V}$. Although the oxide of 70 nm thickness can theoretically hold about 70 V before breaking down, we did not use higher voltages, to avoid too large band-bending at the Si-SiO$_2$ interface. This might have resulted in additional interface-trapped oxide charges (see e.g. Sze [159]) which are not considered in the device simulation.

- The spline coefficients of the boron and phosphorus, as well as the drawn gate length $L_G$ and the phosphorus tail parameters, are finally determined by nonlinear parameter extraction using the optimizer PROFILE.

For both the phosphorus and boron profile, we demand a zero derivative at the high dose boundary of the knot window, allowing the reduction of one spline parameter per profile. Moreover, in the case of phosphorus, the sensitivity for the higher doping levels is too low to allow accurate extraction of the top, and we therefore fix the boundary value as well. For boron, however, the optimizer needs the freedom to scale the profile at the boundary as the simulated vertical profile might differ from the real one due to, e.g., segregation effects, or process spread. Ideally, provided the simulations were accurate and there was no process spread, we should find a boundary spline coefficient which is equal to the simulated top doping level.
and the scaling and spread of the penetrated phosphorus tail.

Table 3.1: The vertical and lateral knot sequences used in the extraction of the lateral doping profile. The position of the knots is actually defined with respect to gate edge as the gate length is also included in the extraction procedure. In this table we assumed that \(L_G = 6 \mu m\), running from 4 to 10 \(\mu m\). Compare with figure 3.13.

<table>
<thead>
<tr>
<th>window</th>
<th>vertical [(\mu m)]</th>
<th>lateral [(\mu m)]</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-type</td>
<td>0.00 0.20 1.20 3.50</td>
<td>2.50 4.00 5.00 7.00</td>
</tr>
<tr>
<td>n-type</td>
<td>0.00 0.38 0.78 1.20</td>
<td>9.10 9.40 9.90 10.5</td>
</tr>
</tbody>
</table>

The results of the parameter extraction are shown in figures 3.12 and 3.13 and in table 3.2. The \(CV\) simulations in figure 3.12 show an excellent fit to the measurements. For comparison, we included a \(CV\) simulation of the default device (\(L_G=6\) \(\mu m\) and no phosphorus reach-through) using rotated profiles and the often-used 'rule of thumb' that the lateral characteristic length is about 0.7 times the vertical one. Some of the parameters turned out to be insensitive and were fixed. The remaining parameters showed good sensitivity and were not coupled, as was found by an eigenvalue analysis of the solution, available within PROFILE [122].

The extracted lateral doping profile is shown in figure 3.13. The spline coefficient at the high concentration boundary of the boron profile turned out to be about 25% lower than the top doping of the simulation \((6.95 \times 10^{17} \text{ cm}^{-3})\), which is, however, just within the extraction accuracy. The phosphorus spline coefficient corresponding to the left-most knot turned out to be insensitive: obviously, we overestimated the phosphorus out-diffusion initially.

Comparison with threshold voltage method

In figure 3.14, we compare the extracted doping profile with the result obtained by the \(V_T\) method given in section 3.3.4. The latter is shifted 50 nm beyond the gate edge to match both profiles. As can be observed, the profiles coincide closely, except near the gate edge and the lateral junction where the 1D assumption of the \(V_T\) method is invalid.
Table 3.2: Extracted spline coefficients describing the lateral profile shown in figure 3.13. The subscripts refer to the knot indices for which the spline coefficients apply. Some parameters turned out to be insensitive and were fixed.

<table>
<thead>
<tr>
<th>parameter</th>
<th>unit</th>
<th>extracted value</th>
<th>remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_1$</td>
<td>cm$^{-3}$</td>
<td>$5.44 \times 10^{17}$</td>
<td></td>
</tr>
<tr>
<td>$p_2$</td>
<td>cm$^{-3}$</td>
<td>$1.92 \times 10^{17}$</td>
<td></td>
</tr>
<tr>
<td>$p_3$</td>
<td>cm$^{-3}$</td>
<td>$1.19 \times 10^{16}$</td>
<td></td>
</tr>
<tr>
<td>$p_4$</td>
<td>cm$^{-3}$</td>
<td>$2.23 \times 10^{14}$</td>
<td></td>
</tr>
<tr>
<td>$n_1$</td>
<td>cm$^{-3}$</td>
<td>$&lt;1 \times 10^{14}$</td>
<td>insensitive, fixed</td>
</tr>
<tr>
<td>$n_2$</td>
<td>cm$^{-3}$</td>
<td>$7.88 \times 10^{15}$</td>
<td></td>
</tr>
<tr>
<td>$n_3$</td>
<td>cm$^{-3}$</td>
<td>$1.76 \times 10^{19}$</td>
<td></td>
</tr>
<tr>
<td>$n_4$</td>
<td>cm$^{-3}$</td>
<td>$9.96 \times 10^{19}$</td>
<td>fixed</td>
</tr>
<tr>
<td>$L_G$</td>
<td>$\mu$m</td>
<td>6.33</td>
<td></td>
</tr>
<tr>
<td>$N_{surf}^{P_{tail}}$</td>
<td>cm$^{-3}$</td>
<td>$1.93 \times 10^{16}$</td>
<td></td>
</tr>
<tr>
<td>$\sigma_y^{P_{tail}}$</td>
<td>$\mu$m</td>
<td>1.0</td>
<td>insensitive, fixed</td>
</tr>
</tbody>
</table>

Figure 3.12: $C_{GB}$ and $C_{GD}$ measurements of the gated diode, together with the simulations corresponding to the extracted profile. The dashed lines were calculated with rotated and laterally shortened doping profiles and default gate length.
Figure 3.13: The extracted lateral doping profile. The high dope of the epi-layer is caused by tail of the SN phosphorus implant which reached through the polysilicon gate and the gate oxide into the epi-layer, thereby increasing the doping level close to the surface.

Figure 3.14: Comparison of the profile measured by the spline method and the $V_T$ method.
3. Determination of doping profiles

Discussion

The spline profiling method can, in principle, be applied to 2D profiles, provided sufficient 2D CV data is available. When using the LDMOST, for instance, it is possible to scan the channel profile by both $C_{GB}(V_G, V_B)$ and the source-back gate capacitance $C_{SB}(V_G, V_B)$. Attempts to include $C_{SB}$ in the optimization for a full 2D extraction of the doping profiles in the channel region failed, because it turned out to be a too difficult a capacitance to model accurately. This can be understood by considering that the out-diffused SN and SP2 profile partly overlap, which gives a large contribution to $C_{SB}$. A more dedicated device design or inclusion of other electrical data, such as drain-back gate breakdown characteristics, in the extraction is then necessary. We did not pursue the latter option as the computational effort would be very large, while it was unclear whether enough additional information could be collected to allow for a full 2D extraction.
Chapter 4

Transient enhanced diffusion

In this chapter, we present experiments of enhanced boron diffusion during annealing of ion implantation damage. This work was initially motivated by the transient enhanced diffusion observed in the SIMS profiles of the DMOS, shown in figure 1.2 in chapter 1.

Transient enhanced diffusion phenomena have become an important issue in IC technology. The effect has been related to anomalous electrical device behavior such as unsuspected threshold voltage shifts in small MOSFETs and low cut-off frequencies in narrow-width bipolar transistors.

Our work was further motivated by the possibility to experiment with high-quality silicon epitaxial layers using the ASM-Epsilon Chemical Vapour Deposition apparatus, which has become operational at DIMES. The low density of defects in this material, in combination with the possibility to grow narrow and steep boron profiles, allows easy detection of enhanced diffusion phenomena, as caused by the damage anneal.

Finally, the temperature of 700 °C in our annealing experiments was such that the results will also be interesting for new and forthcoming IC technologies, such as those based on silicon-germanium, as the thermal budgets keep dropping in an attempt to create very shallow junctions, or to prevent relaxation of strained SiGe layers.

The experiments allow us to set a lower limit for the much disputed silicon and boron interstitial diffusivity at 700 °C in silicon. Boron deactivation has been observed due to silicon self-implantations and implantation damage annealing.
In the following sections, we first give a more detailed introduction to enhanced dopant diffusion. Next, we present a limited introduction to diffusion kinetics in order to understand enhanced diffusion phenomena. Also, several apparent diffusion ‘anomalies’, which are known in the literature, are discussed within the scope of defect and dopant kinetics. Subsequently, we describe our experiments on transient enhanced boron diffusion. In the last section, we return to the DMOST doping profiles and treat the measured vertical doping profiles from a diffusion-kinetics point of view.

4.1 General introduction

With the rapid evolution of silicon IC technology, geared towards making devices faster by reducing their size and using shallower doping profiles, it was soon found that the diffusion behavior of common dopants, such as phosphorus and boron, differed from the basic diffusion laws as defined by Fick. For example, it was observed that high-dose phosphorus profiles exhibits a ‘kink’ and that oxidation enhances diffusion of phosphorus, boron and arsenic, but retarded diffusion of antimony. As it is now well established that dopants diffuse by interaction with native point defects (silicon self-interstitials and vacancies) these diffusion enhancement phenomena are attributed to a—temporary—offset of the equilibrium-native point defect concentrations.

One obvious way to significantly disturb the equilibrium concentration of point defects is by ion implantation. Ion implantation is a common way to introduce a precise amount of dopant in a shallow silicon layer. Heat treatment is required to anneal the implantation damage and to electrically activate the dopant. In order to realize shallow junctions, the thermal budget is kept as low as possible to prevent diffusion of the dopant. However, it was soon observed that during the initial stage of the anneal, the diffusion of the dopant was enhanced. At low thermal budgets, “normal” diffusion is minimal, and the impact of this ‘transient enhanced diffusion’ (TED) is even more pronounced. (In the following, diffusion according to Fick’s law is often referred to as equilibrium diffusion, where ‘equilibrium’ in this case means the absence of diffusion anomalies, i.e., the native point defects are at their equilibrium concentrations.)

The thermal budgets in state of the art technologies are indeed small enough to avoid nearly any equilibrium diffusion, either by low temperature processing, or by very short anneals using rapid thermal annealing. Therefore, TED has become the critical factor in obtaining shallow doping profiles through ion implantation,
4.2. Diffusion

and much effort has recently been spent in understanding, modeling and reducing TED.

When studying the literature, it soon becomes clear that despite more than two decades of research, the kinetics involved in dopant diffusion and damage annealing is still not completely understood, although there has been—and is—a steady progress toward understanding these phenomena. As diffusion of dopants takes place through interaction with native point defects, it is obvious, that for complete understanding, the native point defect parameters, such as the diffusivity and equilibrium concentration, must be known. Unfortunately, due to their low concentrations, they can only be probed indirectly through their interaction with foreign atoms, or with extended defects such as stacking faults. Therefore, the precise diffusivities and equilibrium concentrations of vacancies and silicon self-interstitials are still not known. The controversy about the native point defect parameters is the figurehead of all diffusion-related phenomena, and has long inspired scientists to conduct all kinds of experiments, which resulted in a wide spread of the published point defect parameter data, see, e.g., references 8, 13, 14, 48, 54, 85, 88 and 155.

In order to study TED, the complex kinetics involved calls for dedicated test structures. In this work, we chose to use dopant “markers” to monitor nonequilibrium point defect concentrations during the damage annealing process. In the next section, therefore, we treat some theory to serve as a base to understand (enhanced) dopant diffusion.

4.2 Diffusion

In this section, we present a simplified derivation of the boron diffusion coefficient under nonequilibrium native point defect conditions. This is sufficient to interpret the results of the enhanced diffusion experiments in section 4.4.

It has been well established, that the migration of dopants, residing on substitutional lattice sites, occurs by interaction with native point defects, i.e. vacancies and silicon self-interstitials, see e.g. Pichler et al. for direct experimental evidence for boron [127] or Fahey, Griffin and Plummer [38] and references herein. Therefore, in order to understand dopant diffusion, and in particular the enhanced boron diffusion phenomena, it is necessary to understand the interaction between dopant atoms and native point defects.

In the following, we will first define various point defects which are involved
in dopant diffusion, and write down the reactions through which they are formed. Based on these reactions, we next derive the boron diffusivity under nonequilibrium native point defect and intrinsic doping conditions, i.e., the doping level is below the electron-hole concentration during an anneal. Finally, we discuss some literature on measurements to determine the native point defect equilibrium concentrations and diffusivities. For an in-depth treatment of diffusion kinetics and a complete treatment of diffusion equations, the reader is referred to the excellent review article by Fahey, Griffin and Plummer [38]. An extensive treatment of the theoretical and experimental aspects of point defects in semiconductors can be found in the books by Bourgoin [11] and Lanno [86].

4.2.1 Definition of point defects

Figure 4.1 presents a general overview of point defects in a two-dimensional representation of the Si lattice. A distinction is made between “native point defects” which exist in the pure silicon lattice, and “impurity related point defects”.

A vacancy is defined simply as an empty lattice site, and denoted \( V \). The silicon self-interstitial is a silicon atom which resides in one of the interstices of the lattice, whereas the silicon self-interstitialcy or split Si-interstitial consists of two displaced atoms around one lattice site, and is therefore different from the true tetrahedral interstitial atom. Both defects comprise one extra silicon atom, and from a macroscopic point of view, a distinction cannot be made. Both are therefore referred to as \( I \).

The impurity related defects are defined accordingly. The substitutional dopant is written as \( A_s \), the dopant interstitial as \( A_i \), the dopant interstitialcy as \( AI \), and the dopant-vacancy pair as \( AV \). Only dopant defects other than the substitutional atom can migrate, unless we assume that the substitutional dopant atom and a neighboring silicon atom exchange places spontaneously, which is considered to be highly improbable [38]. Discarding the charge state of the point defects, the following reactions describe their relation:

\[
A_s + V \rightleftharpoons AV, \tag{4.1}
\]

\[
A_s + I \rightleftharpoons AI, \tag{4.2}
\]

\[
A_s + I \rightleftharpoons A_i, \tag{4.3}
\]

\[
A_s \rightleftharpoons A_i + V. \tag{4.4}
\]

Reactions 4.2 and 4.3 are known as the “kick-out” mechanism and represent a substitutional-interstitial(cy) interchange process. The reverse reaction is referred to
as "kick-in". From a macroscopic point of view, the dopant interstitial and dopant interstitialcy are, like the silicon interstitial(cy), indistinguishable. Reaction 4.4 is known as the "dissociative mechanism", or "Frank-Turnbull mechanism" [46] and is a competing process for generating dopant atoms in an interstitial-type state. For boron, it has been found that the kick-out mechanism controls the diffusion [26].

In order to keep formulations short, we often refer to the silicon self-interstitial as "Si-interstitial", whereas the dopant interstitial is always named completely.

### 4.2.2 Boron diffusion equation and effective diffusivity

The first step towards a macroscopic diffusion equation for boron, based on pair diffusion theory, is the identification of the particular pairing reactions underlying boron migration. It is now firmly established that boron, under intrinsic doping conditions, diffuses primarily in an interstitial-type state, see for direct experimental evidence [25,27,128] or Nichols et al. [112] for ab initio calculations. As from a macroscopic point of view, the distinction between dopant interstitial and dopant interstitialcy is not important, we assume, for the sake of simplicity, that the boron interstitial is the migrating diffusing species. We thus have only to consider
the kick-out reactions with the silicon interstitial in various charge states:

\[ B_{s}^{-} + I^{\pm c} \rightleftharpoons B_{i}^{-1\pm c}, \]  \hspace{1cm} (4.5)

where \( B_{s}^{-} \) and \( B_{i}^{-1\pm c} \) are the substitutional boron and interstitial atom respectively. \( c \in \{0,1,2\} \) describes the various charge states of the point defects. It is generally assumed that the number of displaced (and mobile) dopant atoms in equilibrium is much smaller than the number of substitutional (and immobile) atoms, i.e. \( B_{i}^{-1\pm c} \ll B_{s}^{-} \). The macroscopically observed diffusion is thus governed by the fraction and diffusivity of the mobile species. These, in turn, depend on the formation and the migration energy of the dopant interstitial, of which a schematic representation is given in figure 4.2. The impact of an increase of silicon interstitials on boron is now obvious: a larger mobile boron fraction results through reaction 4.5, thereby increasing the macroscopically observed diffusion.

Now that we have presented the microscopic mechanisms of boron diffusion, the next step is to identify those mechanisms that can change the dopant distribution, which are:

- linear diffusion;
- drift by an electric field;
- reaction with other species.

Figure 4.2: Schematic representation of the mechanism of impurity defect migration (here, boron is taken) via a mobile intermediate interstitial species. a) The fraction of mobile species is determined by the formation of an interstitial impurity through, e.g., the kick-out reaction. b) Illustration of the migration energy which is required by the mobile species to move from one interstice to another.
4.2. Diffusion

The three effects can be lumped in a generalized diffusion equation for each species. Obviously, when considering all species in their various charge states, a large set of kinetic reactions and diffusion equations results. In order to solve this set, the diffusivity and reaction rate constants of each species must be known. Although such a set can, in principle, be solved by brute force computing, most problems allow a few approximations thus yielding a strongly reduced set of partial differential equations in which effective diffusivities appear. We therefore make the following assumptions:

- Intrinsic diffusion; the doping concentrations are below the intrinsic electron-hole concentrations during an anneal. Although it is possible to include extrinsic diffusion and electric field effects in the analysis, the equations become more complicated;

- Electronic charging reactions, such as $I^0 + e^- \leftrightarrow I^-$, are in equilibrium.

The concentrations of a specific point defect at different charge states then have a fixed ratio, which allows us to contract the corresponding equations to one diffusion equation. We can now write the following diffusion equations for $B$, $I$ and $V$, where we have dropped the charge states:

\[
\frac{\partial}{\partial t} C_{B_i} = \frac{\partial}{\partial x} \left( D_{B_i} \frac{\partial}{\partial x} C_{B_i} + G \right)
\]

\[
\frac{\partial}{\partial t} C_{B^*} = -G
\]

\[
\frac{\partial}{\partial t} C_I = \frac{\partial}{\partial x} \left( D_I \frac{\partial}{\partial x} C_I \right) - G - K(C_I C_V - C_I^{eq} C_V^{eq}) \quad (4.6)
\]

\[
\frac{\partial}{\partial t} C_V = \frac{\partial}{\partial x} \left( D_V \frac{\partial}{\partial x} C_V \right) - K(C_I C_V - C_I^{eq} C_V^{eq}),
\]

in which $C_X$ denotes the concentration of point defect $X$. $D_{B_i}$, $D_I$ and $D_V$ are the diffusivities of the dopant interstitial(cy), Si-interstitial and vacancy, respectively, $G$ is the generation rate of dopant-interstitial pairs by the kick-out mechanism, $K$ is the bimolecular recombination rate reflecting bulk recombination of Si-interstitials and vacancies which exhibits an activation energy of 0.3 eV [48]. Other sources or sinks of interstitials and vacancies may be added to this system, e.g., generation/recombination at the surface.

To reduce this set further to three equations, one for Si-interstitials, one for vacancies and one for boron, we additionally assume that:
• the pairing reactions 4.5 are in local equilibrium;

• the native point defect gradients in the equation for boron are negligible.

The latter assumption is not essential and it should be be mentioned that it is not met in many practical cases with high excess native point defect concentrations near strong sinks. However, it does apply to the boron enhanced diffusion experiments reported in section 4.4, and it greatly simplifies the analysis.

We proceed by summing the first two equations of 4.6 thus yielding one equation for the total dopant concentration $C_B$:

$$
\frac{\partial}{\partial t} C_B = \frac{\partial}{\partial x} \left( D_{Bi} \frac{\partial}{\partial x} C_{Bi} \right)
$$

$$
C_{Bi} = \gamma C_{B} C_I
$$

$$
C_B = C_{Bi} + C_B
$$

(4.7)

where $\gamma$ is the reaction constant of the kick-out reaction. As $\gamma$ is assumed constant, it can be written in terms of equilibrium concentrations:

$$
\gamma = \frac{C_{Bi}^{eq}}{C_B^{eq}} \frac{1}{C_I^{eq}},
$$

(4.8)

where it is additionally assumed that, at equilibrium, $C_{Bi} \ll C_{Bi}$, so that $C_{Bi} \approx C_B$.

After some rearrangements, we finally find the desired form in $C_B$:

$$
\frac{\partial}{\partial t} C_B = \frac{\partial}{\partial x} \left( D_B \frac{\partial}{\partial x} C_B \right)
$$

(4.9)

with the effective boron diffusion coefficient $D_B$ given by:

$$
\frac{D_B}{D_B^{eq}} = \frac{D_{Bi}^{eq}}{D_B^{eq} + C_I^{eq}}
$$

(4.10)

in which $D_B^{eq}$ is the macroscopically measured effective diffusion coefficient for equilibrium-native point defect concentrations. This equation reflects the enhancement of the diffusivity for a supersaturation of silicon interstitials. $D_B$ increases linearly for relatively small supersaturations, but saturates at the boron interstitial(cy) diffusivity for large supersaturations. This theoretically occurs when all
dopants are displaced to become mobile interstitialcies. In this equation, only $D_B^{eq}$ is well known.

Under equilibrium-native point defect conditions, the equilibrium diffusion coefficient is given by (using 4.7, 4.8 and $C_{Bi} \ll C_B$):

$$D_B^{eq} = \frac{D_{Bi}C_{Bi}^{eq}}{C_B},$$

(4.11)

in which the product $D_{Bi}C_{Bi}^{eq}$ is termed the "transport capacity" or "diffusion capacity" which reflects the supply or drain of dopant $B$. As $D_B^{eq}$ is actually measured from doping profile measurements, the determination of its components, $D_{Bi}$ and $C_{Bi}$, is difficult.

**Numerical simulation tools**

Most currently available process simulation tools solve one equation for each dopant and one for the Si-interstitials and vacancies respectively, see, e.g., references 5, 78, 106, 132 and 166. The set of equations which is solved is more complicated than presented here, in order to include extrinsic diffusion and native point defect gradients. Moreover, additional physical models are being developed and added to the simulators to describe, e.g., dopant clustering phenomena [4, 58, 170] or damage annealing [126] (see also section 4.3). This is either done by adding extra equations to the system, or by an "effective diffusivity" approach [104], in which the effects are incorporated in an adapted diffusivity, so as to avoid the necessity to solve additional equations. Currently, much work is being done on modeling implantation damage annealing [66, 130, 126].

### 4.2.3 The Si-interstitial diffusivity

The accurate modeling of nonequilibrium diffusion phenomena requires the knowledge of point defect diffusivities and equilibrium concentrations. In accordance with the previous section and the results presented in subsequent sections, we focus on the Si-interstitial parameters. As the diffusion capacity $D_I C_I^{eq}$ is known experimentally [13, 156], it is sufficient to determine either $C_I^{eq}$ or $D_I$.

Their measurement, however, is not trivial, as they can only be determined indirectly by interaction of the Si-interstitials with foreign atoms or extended defects, or by self-diffusion experiments using, e.g., the —short-lived— radio tracer $^{31}$Si. In the literature, there is a large discrepancy between measured $D_I$ values.
On the one hand, there is a difference between high and low temperature results. Watkins [173] inferred from electron paramagnetic resonance measurements on boron doped silicon that silicon interstitials were already mobile at 4K. This is consistent with results from Larsen et al. [88], who recently presented evidence for interstitial migration and trapping on impurities at room temperature, by measuring dopant deactivation by means of spreading resistance profiling. However, when extrapolating the Si-interstitial diffusivities derived from impurity diffusion experiments, carried out at high temperatures (>700 °C), the diffusivity at room temperature is found to be virtually zero.

On the other hand, the results from the impurity diffusion experiments themselves show a large mutual difference, and reported values differ up to six orders of magnitude at 800 °C.

The difficulty of measuring the interstitial parameters has inspired scientists to conduct numerous experiments to extract them. Impurity diffusion experiments are very useful, as they provide a very sensitive means of monitoring small offsets in the point defect concentrations through the impurity diffusion behavior. They can be divided roughly in two groups: diffusion experiments using fast metal diffusers, and experiments with dopant marker layers and doping superlattices. The results of these two types of experiments show a wide discrepancy in the derived interstitial diffusivity; \( D_I \) is found to be much higher in the metal studies, especially at relative low annealing temperatures.

In metal diffusion experiments, fast interstitial diffusers such as gold [8, 14, 48, 156], platinum [102, 184], or zinc [13] are diffused into the silicon. By assuming interactions with silicon interstitials, such as equation 4.5, Si-interstitial diffusion coefficients are calculated by analyzing the metal diffusion profiles through, e.g., spreading resistance profiling (SRP). In order to determine \( D_I \) and \( C_{eq}^I \) independently, it is necessary to establish a nonequilibrium and nonstatic point defect distribution during the metal indiffusion.

In doping layer experiments, doping layers are grown and their response to Si-interstitial injection is examined. By using steep and narrow doping profiles, even a small amount of —enhanced— diffusion is easily detected. See figure 4.3 for an illustration. From the depth and time dependence of the diffusion enhancement it is, in theory, possible to extract the Si-interstitial diffusivity. However, the diffusion of common dopants is much slower than the Si-interstitial diffusion. On the time scale of dopant diffusion, this results in stationary Si-interstitial profiles which do not provide any direct information on \( D_I \).
4.2. Diffusion

Figure 4.3: Schematic illustration of marker layer experiments. "equilibrium" diffusion refers to diffusion under equilibrium-native point defect concentrations.

Depending on the purity of the material, trapping defects, such as carbon atoms, may be present and act as Si-interstitial traps which slow down the diffusion front [22, 154]. One then measures a lower —effective— Si-interstitial diffusivity. Griffin and Plummer [55] first proposed Si-interstitial traps to explain the differences in measured diffusivities. Gossmann et al. [52] recently showed that the wide spread in published data can be reconciled by using the concept of trap-limited diffusion [22]. They argued that published values for $D_I$ and $C_I^{eq}$ essentially represent the product $D_I C_I^{eq}$, and that they must be interpreted as a lower and higher limit, respectively, when neglecting the presence of traps.

In summary, only the product of $D_I$ and $C_I^{eq}$ is is well known and described by the Arrhenius relation [139]:

$$D_I C_I^{eq} = 1.5 \times 10^{26} \exp \left( \frac{-4.95}{kT/q} \right)$$  \hspace{1cm} (4.12)

The measurement of its components is still the subject of research.

A natural question which arises is, where freely moving excess Si-interstitials remain if the lattice is allowed to relax to equilibrium, as they are already mobile at cryogenic temperatures? We have already mentioned interstitial traps such as carbon. In the next section, we treat this matter in parallel to the phenomena which
are known to inject Si-interstitials into the silicon substrate.

4.3 Anomalous diffusion effects

As diffusion of dopants occurs through interaction with native point defects, it is obvious from equations 4.6 that whenever the point defect concentrations differ from their equilibrium values, dopant diffusivity will be affected. Several processing conditions are known to establish an offset of the equilibrium-native point defect concentrations. In this section, we focus on those conditions that inject a net of silicon interstitials into the substrate. We treat the corresponding diffusion phenomena from the perspective of nonequilibrium interstitial and vacancy concentrations. In the last subsection, we summarize some literature in which anomalous device characteristics were attributed to enhanced diffusion effects.

4.3.1 Oxidation enhanced diffusion

It has been known for a long time that oxidation enhances the diffusion of some dopants (e.g. boron, phosphorus and arsenic), while retarding the diffusion of others (antimony). From stacking fault [63] and dislocation loop growth experiments [96], it was concluded that, during oxidation, a net flux of silicon interstitials are injected, as these defects contained excess silicon atoms (“extrinsic defects”). Thus a supersaturation of interstitials, and an undersaturation of vacancies, through interstitial-vacancy recombination, is established. These experiments are used to determine the relative contribution of the Si-interstitial and vacancy component of dopant diffusion. From these experiments, it is concluded that, e.g., boron mainly diffuses through interaction with interstitials, whereas antimony does by interaction with vacancies.

Several investigations [34,55,94,161] have been carried out to study the lateral extent of oxidation enhanced diffusion (OED), as this is of importance to dopant redistribution in, e.g., MOSFETs which are made by using a gate re-oxidation step [118]. OED of dopants due to local oxidation of the surface, which is achieved by patterning the surface with a grid of varying periodicity, reveals a much smaller lateral penetration depth of the Si-interstitials. This obviously indicates that the surface efficiently sinks Si-interstitials.
4.3.2 Phosphorus enhanced diffusion

Of all common dopants, phosphorus shows the most peculiar diffusion behavior. High dose concentration profiles, diffused from, e.g., a gaseous source, exhibit a flat plateau at the top, a characteristic ‘kink’ and a diffused tail, see, e.g., Yoshida [180]. Moreover, diffusion of other dopants, especially boron, is enhanced in the presence of a high phosphorus concentration, even if the boron and phosphorus profile are spatially separated, see, e.g., Lee and Willoughby [89].

In an attempt to both understand and model phosphorus diffusion, Fair and Tsai proposed a model based on vacancy assisted diffusion [41] which is illustrated in figure 4.4. It is mentioned here, as this well known model describes the characteristic ‘kink’, and is incorporated in the widely used SUPREM3 simulation program. Fair and Tsai assumed that diffusion in the electron-rich high-dose region was dominated by $P^+V^-$ pairs. Below a certain electron concentration, the double negatively charged vacancies cannot be sustained due to the lack of electrons, and the pairs dissociate. The vacancies then diffuse into the bulk, thereby enhancing diffusion in the phosphorus tail region (through $P^+V^-$ pairs) as well as other dopants.

Although the Fair-Tsai model successfully explains the ‘kink’, it did not predict enough diffusion in the top of the profile to explain the plateau. To this end,
Mathiot and Pfister [104] proposed the physically plausible "percolation" phenomenon: above a certain phosphorus concentration (typically $3 \times 10^{20}$ cm$^{-3}$) a vacancy interacts with several phosphorus atoms, thereby increasing the diffusion capacity of phosphorus.

Various experimental results, however, have led to a reconsideration of the Fair-Tsai model [64]. The results showed that boron, phosphorus and arsenic are enhanced in the vicinity of a high-dose phosphorus profile, whereas antimony diffusion is retarded [37, 60, 107]. Extrinsic stacking fault growth has also been observed underneath the phosphorus [3, 113, 156] diffusion. These phenomena resemble the results for oxidation enhanced diffusion (see previous section) and indicate that Si-interstitials are injected during the drive-in of a high level of phosphorus.

Current consensus is that phosphorus diffuses by pairing with both Si-interstitials and vacancies; at high concentrations the vacancy component dominates, whereas at lower concentrations the interstitial component is more important. In this concept, the enhancement of boron and phosphorus in the tail region is attributed to a supersaturation of Si-interstitials. Various authors [107, 108] were able to qualitatively model the plateau and kink in phosphorus profiles by assuming that phosphorus diffuses by both interstitials and vacancies.

4.3.3 Transient enhanced diffusion

As mentioned in the introduction, it has long been known that implanted dopants show a transient of enhanced diffusion during annealing. These transients were related to annealing of implant damage defects [19, 42, 43], which were observed by, e.g., Transmission Electron Microscopy. Packan and Plummer [124] reported TED of stabilized boron layers during the anneal of nonamorphizing Si implantations. Giles [50] earlier found TED of phosphorus below the amorphization threshold. In order to understand TED, it is therefore necessary to identify the responsible defects and their annealing behavior.

Several types of native defects have been identified in ion-implanted silicon. For brevity, we in general refer to "extended defect" to indicate relatively large and stable defects with an identified structure, i.e., dislocation loops and stacking faults, and to "clusters" to indicate relative small and unstable agglomerations of atoms. Whereas this distinction might seem somewhat artificial (obviously a cluster also has a certain structure), it is useful to assist the discussion on the different time scales of TED which are related to the defect and cluster annealing processes.
4.3. Anomalous diffusion effects

So far, three different forms of transient enhanced diffusion on vastly different time scales and temperatures have been identified in literature, to be addressed in subsequent subsections.

Below the amorphization threshold: very low dose implantations

Cowern et al. [28] studied boron layers in response to anneals at low temperatures (T < 600 °C) and very low dose silicon implants (1×10¹¹ cm⁻²). They observed an ultrafast diffusion transient which they attributed to freely moving Si-interstitials escaping the vacancy-interstitial population which is present in the as-implanted material at room temperature.

Below the amorphization threshold: point defect clusters

During implantations and the initial stages of a subsequent anneal, point defect clusters will rapidly nucleate and grow, as suggested from strain measurements by Servidori et al. [143]. Although the precise kinetics of the nucleation process is still not clear, a simple rule of thumb appears to apply: the amount of Si-interstitials trapped in clusters in the early stages of the anneal is about equal to the implanted dose ("+1 model" [50]). These metastable clusters dissolve during prolonged annealing, thereby releasing the captured excess Si-interstitials which enhance the diffusion of dopants. The clusters thus tend to buffer the transient enhanced diffusion: Angelucci et al. [1] found that the enhancement is nearly independent of the implanted dose and diffusion time, in contrast to the total duration of the transient which directly relates to the buffered amount of Si-interstitials (and thus the implanted dose).

Recently, Eaglesham et al. [36] have identified the dissolving "{113} defect cluster" or "rod-like" defect as the source for enhanced diffusion phenomena after nonamorphizing implants; {113} defects run along <110>-directions and consist of excess Si-interstitials precipitated on {113} planes. In a theoretical treatment, Tan [159] showed that the {113}-defects are an energetically favorable way to accommodate excess Si-interstitials in the silicon lattice. Figure 4.5 shows a TEM picture of a typical {113} defect, taken, with permission, from Eaglesham's article.

By measuring the size of the defects during the dissolution process, Eaglesham found an activation energy of 3.6 eV, in perfect agreement with published values for the diffusion transients, see, e.g., Solmi et al. [150]. This also corresponds well with the value of 3.5 eV found already by Wu, who studied rod-like defects in boron implanted silicon [178]. Diffusion enhancements as high as 1×10⁴
at 670 °C have been reported [154] in relation to this transient.

However, Zhang [182] recently reported enhanced diffusion of boron implanted profiles under nearly equal annealing conditions without the observation of \{113\} defects. This is consistent with the observation that the incorporation of boron or carbon atoms in the silicon lattice suppresses the formation of \{113\} defects and TED [115,149,153,155]. Jones [75] explained these results by suggesting that small "boron interstitial clusters" (BIC) are being formed, rather than \{113\} defects. The dissolution of BICs is, then, thought to be responsible for the observed transient. Although the idea of boron incorporation into small clusters is not new, see, e.g., Wu and Washburn [179], Michel et al. [107] or Huang et al. [65], and not unreasonable, there is no direct evidence.

The injected silicon interstitials either diffuse to the interface or into the bulk, where they are annihilated by traps (e.g., carbon [153], small dislocation loops [179]). Lim et al. [93] recently showed that the surface plays a key role in the annealing of nonamorphizing implant. Wu and Washburn [179] actually observed the growth of small dislocation loops while nearby rod-like defects were shrinking, which indicates that dislocation loops are good sinks during the cluster dissolution.
Above the amorphization threshold

At high implant doses, the top layer of the silicon is amorphized. Amorphized silicon rapidly regrows during annealing (solid phase epitaxy; typically in the order of 0.1 $\mu$m$^{-1}$ at 600 °C). At the previous amorphous-crystalline interface ("a/c interface") end-of-range (EOR) dislocation loops are observed [75]. The EOR dislocation loops divide the structure in two parts, illustrated in figure 4.6: a nearly defect-free top layer, and a damaged deeper layer where point defects may nucleate into clusters. Indeed, Servidori et al. [143] indirectly measured the presence of clusters by X-ray diffraction measurements which revealed compressive strain profiles around and beyond the a/c interface after solid phase epitaxy. During annealing, two transient enhancements have been observed.

First, the point defect clusters, presumably \{113\} defects, dissolve. Dislocation loops appear to be effective sinks for Si-interstitials [17, 42, 95]. Therefore, the top layer is shielded from this transient and only diffusion in the tail region is observed [116, 151]. The dislocation loops are not only effective sinks for Si-interstitials; they also getter dopants [17, 116, 140].

Second, the rapid transient merges into the slower transient of dislocation loop

Figure 4.6: After the rapid regrowth of the amorphous layer, EOR dislocation loops are found on the former a/c interface. During annealing, two transients on different time scales are observed: one which is attributed to dissolving point defect clusters and one which is associated with dissolving dislocation loops.
dissolution, which causes a second enhanced diffusion regime which is longer, but much less strong.

4.3.4 Dopant clustering and precipitation phenomena

In contrast to enhanced diffusion, some dopants exhibit an inactive and immobile peak at high concentrations. In section 4.4.5, we observe boron immobilization and we therefore address some literature on this effect.

Arsenic is known to exhibit an electrically inactive peak, attributed to precipitation or clustering. There is no consensus yet about the actual deactivation mechanism, though recent results and arguments appear to be in favor of a clustering mechanism [135]. It is important to know whether and how native point defects are involved in the clustering mechanism, as this may yet be another source for diffusion enhancements or retardation. Indeed, recent experiments by Rousseau et al. [134,135] show that arsenic deactivation injects Si-interstitials into the bulk, as detected by enhanced diffusion in a buried boron layer.

High dose boron implantations followed by an anneal show an inactive, immobile peak and TED in the tail region, see, e.g., references 65, 107 and [151]. Much like phosphorus, they show a ‘kink’. For relatively low annealing temperatures (<900 °C) the kink coincides with the intrinsic carrier density and at higher temperatures with the solid solubility level. The striking coincidence with the intrinsic carrier concentration suggests that the electron Fermi level played an important role in the activation and diffusion process, and many researchers developed theories and models based on this observation [4,40,61]. Later, experiments with several doping marker layers showed a depth dependence of the kink concentration, the kink being higher at larger distances from the implant damage region [154]. As the intrinsic carrier density is obviously the same throughout the structure, the observed correlation is probably a coincidence.

The deactivated portion of the boron persists for considerably longer times than the transient diffusion [24, 154], suggesting that the boron is trapped in defects which are more stable than the point defect clusters which cause TED during annealing. This led to the proposition that boron clusters are formed. The simplest possible explanation is the pairing of two boron atoms through:

\[ B_s + I \rightarrow B_i \]  \hspace{1cm} (4.13)

\[ B_i + B_s \rightarrow B_i - B_s \]

where \( B_i \) is the boron interstitial(cy) and \( B_i - B_s \) the inactive and immobile boron
pair cluster [106]. Indeed, this formulation was sufficient to describe the results of enhanced boron clustering experiments in reference 154. However, there is no consensus yet about the actual deactivation mechanism, and without further evidence, other explanations are also possible. For instance, boron could be incorporated in, or initiate, silicon clusters, see, e.g., references 65, 74 and 107. In such a clustering model, the following equations may then be favored:

\[
B_s + I \rightleftharpoons B_i \\
mB_i + B_s + nI \rightleftharpoons BIC
\]  

(4.14)

where \(BIC\) is a boron-interstitial cluster and \(m\) and \(n\) are to be determined from measurements. An analysis of a boron implanted profile by Cowern [23] shows that about one silicon interstitial is captured by each deactivated boron atom \((n = 1)\), consistent with the concept of volume compensation: the volume mismatch of the undersized boron atom on a substitutional lattice site is compensated by trapped excess interstitials.

### 4.3.5 Relation to device behavior

The impact of TED on the device performance has long been recognized. In MOSFETs, for instance, the source and drain implants cause a redistribution of the channel doping which, particularly in small devices, changes the threshold voltage. In SiGe heterojunction bipolar transistors, it is important to keep the base doping profiles, inside the strained SiGe region to ensure optimal device performance, see, e.g., Ghani et al. [49] or Xu et al. [180]. Van Dort [35] showed the impact of TED and OED on the electrical characteristics of small bipolar transistors in a BiCMOS process.

In addition to the profile broadening caused by TED, two other effects must be considered in relation to a high offset of the native point defect concentration:

- **dopant deactivation** — enhanced clustering and deactivation of boron is observed in response to damage annealing (see Stolk [154]); previously active boron may be deactivated;

- **'uphill' diffusion** — high gradients of native point defects can exist near effective sinks such as silicon/oxide interface or dislocation loops. During the dissolution of, e.g., \(\{113\}\) defects, these gradients can be maintained for some time and act as a driving force for dopant diffusion. In the derivation of equation 4.9 from equation 4.7, we neglected the Si-interstitial gradient.
These effects will influence device behavior. For instance, in small NMOS-transistors, anomalous threshold voltage behavior is attributed to uphill diffusion of boron in the channel region (Rafferty [130]). Orlovski [117] reported dopant distribution in the channel region after a gate reoxidation step. Perng and Lin [126] report improved short-channel effects due to oxidation-induced boron redistribution for counter-implantation p-type MOSFETs, thereby actually employing the dopant redistribution. Sadana et al. report enhanced threshold voltage behavior for small NMOSFETs as a consequence of source/drain implant damage annealing. Hanafi et al. [57] additionally describe a model to model anomalous threshold voltage behavior of NMOSFETs based on dopant redistribution.

As thermal budgets will keep dropping in forthcoming technologies, anomalous diffusion effects will play a key role in determining doping profiles and therefore a thorough understanding of implant damage annealing is necessary in order to assist the development of next generation devices.

4.4 Implantation damage annealing and enhanced boron diffusion experiments

In this section, we present new measurements of implantation damage annealing induced enhanced boron diffusion. The injection of interstitials during the anneal of nonamorphizing silicon implants has been monitored by means of sharp boron-doped marker layers grown by reduced pressure chemical vapor deposition. We report an ultrafast diffusion pulse during the initial annealing stages (t < 15 seconds) at 700 °C. The boron diffusivity enhancement is at least an order of magnitude larger than the enhancement during subsequent annealing, cf. figure 1.4. The ultrafast diffusion sets lower limits for the silicon and boron interstitial diffusivities at 700 °C of $2 \times 10^{-10}$ cm$^2$s$^{-1}$ and $2 \times 10^{-13}$ cm$^2$s$^{-1}$, respectively. During the anneals, part of the boron profiles remain stationary, consistent with interstitial-driven boron clustering. Moreover, from SRP measurements, implantation damage induced boron deactivation is found. An analysis combining the SIMS and SRP results shows that these defects, presumably small clusters, are less stable than the defects in which boron resides during annealing.

In the following subsections, we first describe the measurement strategy and the test structures. Next, we show measurement results for an initial exploratory batch of experiments to assess the intrinsic quality of the grown layers. The results led to the setting up of a subsequent batch, of which, in a following subsec-
tion, we present the results on fast interstitial injection during damage annealing. The second-last subsection treats boron deactivation, both induced by the silicon implant, as well as during the anneal. We conclude with a discussion in the last subsection.

4.4.1 How to study damage annealing effects?

Various methods can be used to study defects in silicon. Deep level transient spectroscopy (DLTS) is generally used to detect the presence of specific microscopic defects such as divacancies. X-ray diffraction measurements may be used to study the strain in the lattice, which is used by Servidori et al. [144] to corroborate the presence of clusters. After annealing, cross-section and planar transmission electron microscopy (TEM), or high resolution electron microscopy (HREM) reveal extended defects in and near the implanted surface layers. TEM can also be used to study their size and concentration in order to gain insight in their annealing behavior, see, e.g., references 36, 75 and 83.

Unfortunately, these, and other currently available measurement techniques, cannot observe small clusters directly, nor do they reveal the interstitial supersaturation which is present during implantation damage annealing. For this purpose, doping profile marker layers, or “δ-layers”, as discussed in section 4.2.3, may be grown. The damage annealing process is then monitored by probing the accompanying interstitial supersaturation $C_I/C_{I}^{eq}$ (cf. equation 4.10) through the enhanced diffusion in the marker layers. The profiles can be analyzed using by SIMS and/or SRP analysis techniques.

In the marker layers, we require a dopant which predominantly diffuses by interaction with interstitials to ensure optimal sensitivity. Of the common dopants, boron is the best choice:

- Boron is believed to diffuse primarily through interaction with silicon interstitials and shows the highest diffusion enhancement, together with phosphorus, of all common dopants in response to a supersaturation of Si-interstitials.

- The incorporation of narrow boron profiles during epitaxial growth is very well controlled. This enables us to grow very narrow doping markers with a precise boron content.

- Gossmann et al. concluded from oxidation enhanced diffusion experiments
that the boron marker layers themselves do not influence the silicon interstitial redistribution. Although this remains to be verified in implant damage experiments, where a much larger $C_I/C_I^{eq}$ must be considered, it further stimulates the use of boron $\delta$-layers as "$C_I/C_I^{eq}$ probes".

- Moreover, the range of boron concentrations $(1 \times 10^{17} - 1 \times 10^{19} \text{ cm}^{-3})$ used in these types of experiments is about equal to typical base concentrations in (heterojunction) bipolar transistors and channel regions in MOSFETs. This permits the direct use of the present results to understand the profile behavior in actual devices.

The combination of narrow and steep boron profiles and the high—enhanced—diffusivity allows profile broadening to be detected easily.

Often, silicon is implanted to create the implant damage, since there is a fairly accurate estimate for the amount of excess interstitials that form during annealing. In addition, chemical artifacts in the diffusion processes are avoided.

### 4.4.2 Test structures and measurement methods

#### Test structures

Technology has now evolved to the point that we can grow very pure epitaxial silicon containing very thin boron-doped layers, and anneal this material for times as short as 5 seconds using rapid thermal annealing (RTA). In the present study, eight or nine boron doping spikes of 10, 15 or 20 nm width were incorporated into epitaxially grown Si layers at 0.1 $\mu$m intervals using an ASM-Epsilon reduced pressure chemical vapor deposition (RPCVD) reactor. The deposition temperature was 750 $^\circ$C, the $\text{B}_2\text{H}_6$ partial pressure was $4 \times 10^{-5}$ Torr and the total pressure 60 Torr, which yielded a typical deposition rate of 20 nm/min. The residence time of the gases in this reactor is very short and the thermal variations are small, which enables the growth of sharp doping profiles. Figure 4.7 is a schematic illustration of the structures which are summarized in table 4.1.

In an initial exploratory batch ("A"), we grew eight spikes at 0.1 $\mu$m intervals, seven of which had a top boron concentration of $2 \times 10^{18} \text{ cm}^{-3}$ and one of $1 \times 10^{19} \text{ cm}^{-3}$. The spike width was 10 nm for the low and 15 nm for the high spike. The low spikes were meant to monitor the interstitial supersaturation, whereas the high one was included to observe boron clustering which might affect the interstitial concentration through reactions 4.14. The position of the high spike was varied between 0.1 and 0.5 $\mu$m, resulting in four different structures. In a following, more
Figure 4.7: Schematic presentation of the test structures of the batch 'A' and batch 'B'. The samples with a high spike are shown as a variation on the reference samples without a high spike (A3 and B1). In batch A, 8 spikes were grown at 0.1 μm intervals. In the reference sample A3, all spikes had a top doping level of $2 \times 10^{18}$ cm$^{-3}$ and a width of 10 nm. In the other samples, one of the spikes had a top doping of $1 \times 10^{19}$ cm$^{-3}$ and a width of 15 nm. The position of the high spike is varied to study its influence on the interstitial diffusion. In batch B, an extra spike in the surface region was included. The top doping level was $5 \times 10^{17}$ cm$^{-3}$ for the low spikes and $4 \times 10^{18}$ cm$^{-3}$ for the high one, whose position was in this case fixed at 0.5 μm depth. In this case (batch B) all the spike were 20 nm wide.

extensive batch ('B'), we grew two different structures based on the results from batch A. In an attempt to minimize the influence of the boron spikes on the in-
Table 4.1: Overview of the experiments carried out. 'A' and 'B' refer to the initial and subsequent batch, '1', '2', '3', and '4' denote the structure variation. 'H.S.' indicates the position of the high spike. In figure 4.7, a sketch of the corresponding structures is given.

<table>
<thead>
<tr>
<th>batch</th>
<th>H.S. [μm]</th>
<th>15&quot;</th>
<th>2'</th>
<th>15'</th>
<th>40'</th>
<th>60'</th>
<th>Si dose [cm⁻²]</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>0.2</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>none or 5×10¹³</td>
</tr>
<tr>
<td>A2</td>
<td>0.4</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5×10¹³</td>
</tr>
<tr>
<td>A3</td>
<td>-</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5×10¹³</td>
</tr>
<tr>
<td>A4</td>
<td>0.5</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5×10¹³</td>
</tr>
<tr>
<td>B1</td>
<td>-</td>
<td>*</td>
<td>*</td>
<td></td>
<td>*</td>
<td></td>
<td>2×10¹³ or 5×10¹³</td>
</tr>
<tr>
<td>B2</td>
<td>0.5</td>
<td>*</td>
<td>*</td>
<td></td>
<td>*</td>
<td></td>
<td>2×10¹³ or 5×10¹³</td>
</tr>
</tbody>
</table>

Interstitial indiffusion, the boron peak concentration in the first structure (B1) was chosen to be 5×10¹⁷ cm⁻³, which is still sufficiently high to enable characterization of the atomic profiles by SIMS. In the second structure (B2), a higher spike at 0.5 μm depth, which had a top concentration of 4×10¹⁸ cm⁻³, was grown. In both structures, the spike width was 20 nm.

To control the amount of excess silicon interstitials, silicon was implanted at 40 keV with doses of 5×10¹³ cm⁻² in batch A, and 2×10¹³ cm⁻² and 5×10¹³ cm⁻² in batch B. The wafers of batch A were annealed in a furnace at 700 °C for 15 minutes or 60 minutes. Annealing of the various samples in batch B was done in an AG 610i RTA system for 15 seconds or 2 minutes, or in a furnace for 40 minutes, at 700 °C. Care was taken to avoid temperature overshoot in the RTA system by direct optimization of the pyrometer-based RTA control loop using a thermocouple mounted on a dummy silicon wafer.

**Measurement methods**

SIMS measurements were carried out with a CAMECA IMS 4f instrument using a focused 3 keV, 500 nA O₂⁺ primary beam. The sputtered area was 250×250 μm, of which the central area of 30 μm diameter was used for the analysis of secondary ions. The uncertainties in the concentration and depth scales are estimated to be 10% and 5%, respectively. Two SIMS measurements (A3+A4) were done by Evans Europa.

The SRP measurements were partly carried out by Semiconductor Assessment Services and partly by Privitera at CNR–IMETEM, Catania, Italy. The annealed
samples measured by Privitera were first covered with approximately 0.5 μm silicon oxide, deposited at 350 °C, to allow absolute positioning of the depth axis with respect to the surface. The uncertainties in depth scale and dose are estimated to be 2% and 30% respectively.

4.4.3 Initial experiments: clustering and enhanced diffusion

Batch A was set up in analogy to the multilayer experiments presented by Van Oostrum et al. [116] and Stolk et al. [154]. Stolk used molecular beam epitaxially grown silicon exhibiting a relative large concentration of Si-interstitial traps in the form of a high background concentration of carbon (≈ 1×10^{18} \text{ cm}^{-3}), thereby slowing down the interstitial diffusion front. This resulted in a depth dependence of the measured boron diffusion enhancement. Van Oostrum used atmospheric pressure chemical vapor deposited silicon, which hardly showed any depth dependence of the diffusion enhancement. The purpose of batch A was to get an initial idea about the intrinsic quality of the grown silicon by examining the depth dependence, amount and duration of the diffusion transient.

Figure 4.8a shows the SIMS results for the as-grown and 60 minutes annealed profiles of structure A1, both with and without an Si implant of 5×10^{13} \text{ cm}^{-2}. As can be observed, some migration of boron is found for the unimplanted sample. However, when silicon was implanted we found the profiles A1, and A2 shown in figure 4.8b, to be entirely smeared out, except for the high spikes which are partly immobile and considered clustered (consistent with the result of Stolk et al. [154]). If we describe the high spike by using two Gaussians, one for the mobile part and one for the clustered part, we find clustered fractions of 0.4 and 0.17 for structure A1 and A2, respectively. The interval between the spikes is approximately 10–15% less than intended.

Samples A3 and A4 given in figure 4.9 were annealed for 15 minutes and they clearly show less diffusion. The individual spikes can still be analyzed, except for the spikes neighboring the high one: they disappear in the mobile part of the high one.

An interesting observation from derived this measurement is the clear difference between the spike shape before and beyond the high spike. This strongly suggests that the shallower spikes are also partly clustered. It complicates the use of the low spikes as $C_I/C_{eq}^I$ probes. The clustered fraction of the high concentration marker layer is estimated to be 0.08.

The diffusion behavior of the corresponding spikes in A3 and A4 do not differ significantly, which indicates that the presence of the high spike did not affect the
Figure 4.8: SIMS measurements of structures A1 and A2. Initial experiments which show the response of sharp as-grown spikes to a $5 \times 10^{13}$ cm$^{-2}$ silicon implant annealed for 60 minutes in a furnace at 700°C. Of A1 three profiles were measured: as-grown, no implant + anneal, and implant + anneal. b) For A2 we only measured the annealed sample. Note that the high spikes have in part remained stationary, indicating enhanced boron clustering.

interstitial distribution, i.e., the boron immobilization did not consume considerable amounts of silicon interstitials.

To summarize, we conclude from these results that:

- the silicon we used was of very good quality with relative few interstitial
Figure 4.9: SIMS measurements of structures A3 and A4. These samples were annealed 15 minutes under conditions otherwise the same as structures A1 and A2 in figure 4.8. We note the difference in activation level of the low spikes in the surface region and the high spike (see text).

traps such as carbon, as the enhanced diffusion is observed throughout the entire structure;

- enhanced clustering is observed up to depths of 0.5 μm, even for the lowly doped spikes. This complicates the analysis of the diffusion enhancements to extract values for the interstitial supersaturation \( C_I/C_I^{eq} \);

- the influence of the high spike on the diffusion behavior of the other spikes is not significant, which indicates that the boron immobilization does not consume large amounts of silicon atoms;

- shorter annealing times are required to reveal the initial stages of damage annealing.

These observations led us to set up batch B, where the top doping levels were lower and the annealing times were reduced to time periods as short as 15 seconds.
4.4.4 Ultrafast injection of silicon interstitials

Eaglesham et al. [36] have claimed that TED, in the absence of dislocation loops, can be fully explained by interstitials emitted from dissolving \{113\} defects. In this section, we present evidence for ultrafast TED occurring on a time scale much shorter than the \{113\} dissolution process, even for implant doses at which \{113\} defects are known to be formed. This initial enhancement is strong enough to influence the profile even after prolonged diffusion, and must therefore be considered when evaluating the effects of TED on dopant distributions after low thermal budget annealing.

**Measurements**

Figures 4.10 through 4.13 show the measured doping profiles of batch B after implantation and after various annealing steps. The broadening of the boron spikes is well in excess of equilibrium diffusion, this reflects the enhanced boron diffusion caused by the injected interstitials. In particular, the boron diffusivity during the first 15 seconds appears to be considerably higher than during subsequent annealing. Most of the first dopant spike becomes immobile within 15 seconds, and this immobile component persists noticeably up to 40 minutes. Stolk et al. [154] have shown that a supersaturation of interstitials is responsible for boron deactivation outside the damage implant region. They observed boron deactivation down to concentrations of $1 \times 10^{18}$ cm$^{-3}$ in molecular-beam grown epitaxial silicon films. However, in the present experiment, boron deactivation occurred to much lower concentrations. This probably reflects the much higher supersaturation of interstitials present in the very early stages of annealing as a result of a lower level of Si-interstitial trapping centers (i.e. carbon) in the presently studied marker layers.
Figure 4.10: SIMS profiles on structure B1 after a $2 \times 10^{13}$ cm$^{-2}$ Si implantation and annealing at 700 °C. For a better comparison, the profiles are shown in time-consecutive pairs: a) no anneal and 15 seconds RTA. b) 15 seconds and 2 minutes RTA. c) 2 minutes RTA and 40 minutes furnace anneal. In each case, the initial profile is represented by a solid line and the final profile by symbols.
Figure 4.11: SIMS profiles on structure B1 ($5 \times 10^{13} \text{ cm}^{-2}$ Si implantation; 700°C anneals).
Figure 4.12: SIMS profiles on structure B2 \((2 \times 10^{13} \text{ cm}^{-2})\) Si implantation; 700 °C anneals.
Figure 4.13: SIMS profiles on structure B2 ($5 \times 10^{13}$ cm$^{-2}$ Si implantation; 700 °C anneals).
Analysis

The average boron diffusivity per time interval has been extracted from the profiles in figures 4.10–4.13 using the SUPREM3 simulation program. Only the four deepest spikes are used for this analysis, since for the shallower ones it is not possible to distinguish clearly between the mobile and immobile parts. An enhanced diffusivity is fitted during the three periods (0–15 seconds, 15 seconds–2 minutes and 2–40 minutes) by taking the SIMS profile at the start of each time period as input to the corresponding SUPREM3 simulation. This procedure enables us to distinguish the effect of the ultrafast pulse from the longer TED enhancement. The diffusivity is extracted for each individual marker layer and the enhancement factor is calculated relative to the boron diffusivity under equilibrium point defect concentrations, $D_B^{eq} = 9.1 \times 10^{-19}$ cm$^2$s$^{-1}$ at 700 °C [39]. For the 40 minutes anneal, the individual spikes can no longer be resolved, and only the out-diffusion of the deepest one is analyzed.

Figure 4.14 shows the time evolution of the enhanced diffusivity of this spike for two implantation doses and both structure B1 and B2, and figure 4.15 summarizes the results of this analysis for the individual marker layers. The average dif-
fusivity enhancement is about $2 \times 10^5$ during the first 15 seconds anneal and drops to $1 \times 10^4$ in the period 15 seconds – 2 minutes. Interestingly, the dependence of TED on the implant dose is very weak in the early annealing stages, and only becomes visible after longer annealing times. This point will be discussed in more detail later in this section.

![Graph](image)

Figure 4.15: Depth dependence of the boron diffusivity enhancement calculated from profile broadening in figures 4.10 through 4.13. From structures B2, figure b), the high spike could also be analyzed, although with less accuracy, due to the clustering.

We now interpret the experimental results, starting with the ultrafast pulse. We
recall the following expression for the boron diffusivity, appropriate for the case of interstitial-mediated diffusion under intrinsic doping conditions:

$$\frac{D_B}{D_B^{eq}} = \frac{D_{Bi}^{eq} C_i^{eq}}{D_{Bi}^{eq} C_i^{eq} + \frac{D_{Bi}^{eq}}{C_i^{eq}}}$$

(4.15)

Here, again, $D_B$ is the boron diffusivity, $D_{Bi}$ the boron interstitial(cy) diffusivity, $C_i$ the interstitial concentration, and the superscript "eq" indicates values of $D_B$ and $C_i$ under equilibrium point defect conditions.

Assuming that the enhancement is saturated (i.e., all boron atoms have been displaced and have become mobile) we calculate a lower limit of $D_{Bi} = 2 \times 10^5 \times D_B^{eq} \simeq 2 \times 10^{-13} \text{ cm}^2\text{s}^{-1}$ for the boron interstitial(cy) diffusivity. This value coincides quite closely with the lower limit of $5 \times 10^{-13} \text{ cm}^2\text{s}^{-1}$ estimated by Schork et al. from radiation-enhanced diffusion experiments [138]. Since we observe a depth dependence of the enhancement in our experiments, we believe that the enhancement has not yet saturated and that the data reflect the actual averaged interstitial supersaturation profile $C_i/C_i^{eq}$. The depth dependence, however, is small, even for the 15 seconds anneal, suggesting that Si-interstitial indiffusion must have occurred fast.

A lower limit for the Si-interstitial diffusivity can be calculated by assuming that the front of these diffusing interstitials has reached the last spike within 15 seconds. From SRP measurements on as-implanted samples (see next section) the maximum penetration depth of implant damage is found to be 0.4 μm, so we may conclude that the excess Si-interstitials have diffused over a distance of at least 0.6 μm in 15 seconds. This corresponds to a lower limit of $2 \times 10^{-10} \text{ cm}^2\text{s}^{-1}$ for the Si-interstitial diffusivity at 700 °C. The exact value is most certainly larger, since the average interstitial profile is rather flat, indicating that these interstitials have diffused far beyond the deepest boron spike. Recent experiments using a similar approach [153] gave much lower apparent values for the Si-interstitial diffusivity. This discrepancy appears to be a consequence of the relatively high level of carbon trapping centers in the material used in reference [153]. Our lower limit is consistent with metal diffusion studies, e.g., the recent zinc diffusion results of Bracht et al. [13]. Further experiments are needed to nail down the temperature dependence of interstitial diffusion in high-quality CVD silicon.

For the highest implant dose of $5 \times 10^{13} \text{ cm}^{-2}$, the diffusivity enhancement in the time range 15 seconds to 40 minutes remains virtually constant at a value of $1 \times 10^4$, suggesting that, in this time frame, the interstitial supersaturation is buffered by $\{113\}$ interstitial clusters. The measured enhancement factor of $1 \times 10^4$
matches the predictions of the recently developed model for interstitial clustering and dissolution by Rafferty et al. [129]. For the lower dose of $2 \times 10^{13}$ cm$^{-2}$, the same enhancement is found after 2 minutes, but a somewhat lower enhancement is found after 40 minutes. This result suggests that the TED enhancement is nearly independent of dose during $\{113\}$ defect dissolution, but that at the lower implant dose, the reserve of available $\{113\}$ defects becomes depleted during the 40 minutes anneal. A similar dose-independence of the TED enhancement factor, and dose-dependence of the TED time scale, has been reported by Angelucci et al. [1].

Discussion

The boron diffusion enhancement during the ultrafast pulse is well in excess of the enhancement caused by $\{113\}$ dissolution, suggesting a different source for the interstitials. During the initial stages when nucleation of $\{113\}$ defects takes place, the supersaturation of Si-interstitials will be very high. It is not unreasonable to assume that a small fraction of these excess interstitials escape capture by $\{113\}$ defects and diffuse into the structure, causing the observed ultrafast diffusion pulse. It is also possible that the ultrafast pulse is itself buffered by submicroscopic defects which are less stable than $\{113\}$ defects. In either case, the ultrafast pulse merges into the slower TED phase as the absorption and emission rates of Si-interstitials to and from $\{113\}$ defects come into balance. The present results provide a strong reference for testing atomistic simulations (see e.g., Jaraiz et al. [72]) which are designed to model the defect kinetics of the entire implantation and annealing process.

Summary

In summary, we have presented new measurements showing ultrafast Si-interstitial injection from implantation damage in RPCVD epitaxial silicon preceding the defect-controlled TED stage. During short annealing times ($<15$ seconds at 700 °C), a high supersaturation of Si-interstitials is present, which causes a very high boron diffusivity. The high supersaturation is further confirmed by the observation of strongly enhanced clustering of boron in the near surface region to concentrations as low as $\approx 1 \times 10^{17}$ cm$^{-3}$. After this ultrafast pulse, the diffusion enhancement settles down to a constant value, consistent with a steady equilibrium between the Si-interstitial supersaturation and dissolving $\{113\}$ defects. The measurements allow us to extract a lower limit for the Si-interstitial diffusivity in silicon of $2 \times 10^{-10}$
cm$^2$s$^{-1}$, and a lower limit for the boron interstitial(cy) diffusivity in silicon of $2 \times 10^{-13}$ cm$^2$s$^{-1}$ at 700 °C [71].

4.4.5 Boron deactivation

To study the activated portion of the boron spikes during annealing, several spreading resistance profiling measurements (SRP) were carried out. SRP on these types of structures is a challenge, as the steep profiles and structure depth of about 1 μm requires superb analysis. The SRP measurements were done on as-implanted samples ($5 \times 10^{13}$ cm$^{-2}$) of structures B1 and B2 and on annealed samples of structure B1. After a short description of the measurements, we first discuss the results on the as-implanted samples followed by the annealed samples.

Figure 4.16 shows SRP measurements on the as-implanted structures B1 and B2. The first measurement is done by Pawlik from Semiconductor Assessment Services. The profile agrees quite well with the SIMS profile (cf. figure 4.13) for the deepest five spikes, but the shallower spikes are not resolved. The second measurement, done by Privitera from Catania, shows qualitatively the same behavior: only the deepest spikes are resolved. These observations indicate that the ion implantation has resulted in substantial deactivation of the incorporated boron to a depth of approximately 0.4 μm. In figure 4.17, the same profiles are shown, together with the SIMS measurements. For a qualitative comparison, the SRP profiles are scaled to correct for the discrepancy in the SRP and SIMS depth scales, cf. figure 4.16. To match the tops of the SIMS, the SRP depth scales are stretched by 8% and 6% for figures 4.17a and 4.17b, respectively. To keep the the dose constant, the concentrations are reduced accordingly.

Figure 4.18 shows the active profiles after annealing samples from batch B1 for 15 seconds and 2 minutes. Although the measurements do not allow a precise quantitative analysis, it can be observed that most of the initially deactivated boron was reactivated during the 15 seconds anneal, only the first boron spike is inactive.

Implantation induced deactivation

The deactivation of the as-grown boron spikes is caused by of mobile self-interstitials generated by the silicon implant, as was also found by Larsen et al. [87]. They observed deactivation of dopants up to depths of 2.5 μm in lowly doped epitaxial layers following low energy ion implants. Boron atoms are deactivated because Si atoms interact with boron to form electrically inactive pairs or clusters of atoms.
Figure 4.16: SRP measurements on room temperature silicon implanted boron multilayers: a) structure B1 (without the high spike) and b) structure B2 (with the high spike).

Larsen et al. assume that most of the ion-beam generated damage quickly condenses into small immobile defects, and that a small part of the silicon interstitials diffuses into the bulk, thereby causing dopant deactivation. It is important to note that the deactivation process consumes silicon interstitials which must be released if the boron is reactivated again during subsequent annealing.

Despite the deactivation of the near-surface part of the boron superlattice, there is no systematic variation in the width of the spikes in the atomic profiles as a function of position within the superlattice. This observation is consistent with earlier studies which showed no diffusion of Si-implanted boron marker layers at room temperature [27].
Figure 4.17: SIMS and SRP profiles of (a) structure $B_1$ (without the higher spike) and (b) structure $B_2$ (with the high spike) measured after a $5 \times 10^{13}$ cm$^{-2}$ silicon implantation.

The absence of diffusion provides useful information on the type of defects in which the deactivated boron resides. The peak boron concentration in the superlattice is about $5 \times 10^{17}$ cm$^{-3}$, which corresponds to a distance between the boron atoms of approximately 12.5 nm. Thus, defects containing two or more boron atoms can only form if one or more boron atoms migrates a distance of at least 15 nm. The average characteristic diffusion length of boron to form the most simple cluster containing two boron atoms, the $B_i - B_s$ cluster, can be calculated from the following relation, which is valid for trap limited diffusion in a constant background concentration of traps [22]. Here we estimate the trap concentration $N_T$ to
be equal to the top concentration of the spikes in order to obtain a lower limit.

\[ L = \frac{1}{\sqrt{4\pi a_T N_T}}, \]  \hspace{1cm} (4.16)

where \( a_T \) is the trapping radius, assumed to be equal to the interatomic distance (0.235 nm). We then calculate \( L \) to be \( \approx 25-30 \) nm. This distance becomes even larger when considering the wings of the boron markers where the boron concentration is much lower. The concentration of competing boron trapping sites, such as carbon, is very low as it is known that concentrations comparable to that of our boron layers slow down the Si-interstitial indiffusion considerably during annealing [153], which we did not observe. If boron were to trap only at these defects, even larger migration lengths would be calculated. Since we do not observe boron migration at room temperature, we conclude that the deactivated boron is located in defects containing only one boron atom.

Whereas in the damaged surface layer many defects are present which may incorporate a boron atom, and thereby deactivate it, boron deactivation outside this layer is assumed to be caused by the formation of small clusters through reactions such as 4.13 and 4.14 [87, 153]. Kinoshita et al. [84] proposed simply the boron interstitial(cy) to explain the differences they found between boron SIMS and SRP data. However, they discard the results of Watkins who has shown that the boron interstitial is unstable at room temperature [172]. Yet, the formation of clusters
with more than one boron atom, e.g., $B_t = B_s$ clusters, would require migration of the boron atoms over appreciable distances. In the absence of measurable boron migration we therefore assume, as a ‘working hypothesis’, that small clusters have been formed around the boron atoms which act as ‘seeds’. The boron interstitialcy, as proposed by Kinoshita [85], would, in this concept, be just the most simple cluster possible. In this context it is also worth noting that Jones recently proposed the formation and dissolution of “boron interstitial clusters” to explain TED in boron implanted samples in the absence of $\{113\}$ clusters [74, 182].

Deactivation during annealing

We now turn to the results obtained after annealing. Figure 4.18 shows SRP profiles for a $5 \times 10^{13}$ cm$^{-2}$ implanted sample after annealing for 15 seconds and 2 minutes at 700 °C. A significant fraction of the deactivated boron, seen in figure 4.16, has been reactivated during the 15 seconds anneal. This rapid reactivation implies that the initial B-I type defects are easily able to reconfigure such that the boron contained in them reaches a substitutional site. Their break up releases the bound interstitials which may well be the source for the ultrafast diffusion transient we reported in the previous section. Only the shallow spike a depth of about 80 nm, within the implant damage region, remains electrically inactive after these anneals.

The shallowest spike (figures 4.10 – 4.13), however, has diffused noticeably during the first 15 seconds, and has remained stationary for at least 2 minutes. This can be described by the characteristic lengths, shown in figure 4.19, which were determined from Gaussian fits to the measurements. The characteristic lengths were fitted with the optimizer PROFILE [122, 124] which also features a statistical analysis of the extracted parameters. The error thus found (3σ-value) varied between 5-10%. From the figure, it appears that the first spikes have diffused during the first 15 seconds before they became immobile. As the SRP measurements show that the first spike was inactive for both as-implanted and annealed structures, we conclude that the trapped boron must have made a transition from a less to a more stable defect. Under our experimental conditions, defects are only observed in the surface layer (see reference 36) to a depth of approximately 0.1 μm. In this region, the chance that boron is trapped in relatively stable defects is the highest which probably explains the difference in deactivation between the first and deeper spikes during the anneal. At higher implant doses, more defects are formed, thus quenching boron diffusion more efficiently. This explains the dose dependence observed in figure 4.19.
Figure 4.19: The time dependence of the characteristic lengths of the first δ-layer as obtained by fitting Gaussians on the SIMS measurements of figures 4.10 – 4.13.

Figure 4.20: The time dependence of the characteristic lengths of the first and last δ-layer as obtained by fitting Gaussians on the SIMS measurements of figures 4.10 and 4.11 ('low spikes only' samples).

In figure 4.20, the characteristic lengths of the shallowest and the deepest spikes are shown for structures B1. Clearly, the deepest spike diffuses more than the shallow one which remains stationary between 15 seconds and 2 minutes. As the first spike has diffused after 40 minutes (compare also figures 4.10 – 4.13), it
is tempting to suggest that most boron is incorporated in \{113\} defects which anneal out on the same time scale. However, other trapping sites in the near-surface region cannot be excluded. During implantation of Si\(^{28}\), small amounts of N\(_2\) and CO\(_2\), which have the same mass as the implanted silicon atoms, may have been co-implanted [141]. We could not take specific precautions to prevent the co-implantation of these molecules. Additionally, oxygen atoms may be knocked into the silicon, as we implant through a native oxide layer. Fan and Jaccodine [44] observed a reduction of TED for through-oxide boron implants, which they attributed to gettering of interstitials by recoiled oxygen in the damaged region, thereby referencing to TEM studies on oxide precipitates by Maher et al. [99] and Freeland [47].

4.4.6 Discussion

Our results, show that, at room temperature, boron is deactivated by mobile interstitials generated by the ion implantation process. The absence of boron diffusion, even at very low boron concentrations, indicates that, at room temperature, the deactivated boron resides in defect complexes which contain only one boron atom. The number of deactivated atoms in our experiment is about one order of magnitude smaller than the number of implanted Si ions (approximately \(4 \times 10^{12} \text{ cm}^{-2}\)). If, as estimated from recent experimental studies, each boron atom traps about one self-interstitial [23], then we may infer that about one-tenth of the implanted \(+1\) interstitials are trapped by boron atoms. As most of the deactivated boron becomes reactivated during the initial stages of annealing (within 15 seconds at 700°C), this suggests that roughly one tenth of the inventory of \(+1\) interstitials will be released by the boron during the initial phase of annealing.

The ultrafast diffusion pulse of boron diffusion that occurs in the initial stages of thermal annealing, as reported in the previous section, is well in excess of the enhancement caused by \{113\} dissolution. This is further evidence for a source of interstitials other than \{113\} defects. The mean square diffusion length \((2Dt)\) after the initial ultrafast pulse is about one tenth of the mean square diffusion length after the slower \{113\}-based transient. Again, this is expected if one tenth of the inventory of \(+1\) interstitials is released during the activation of boron. It is therefore tempting to associate the ultrafast transient with the boron activation process. Recent experimental results, which show no ultrafast transient when Si is implanted into a boron-free capping layer [18] are also consistent with this view.

Figure 4.21 summarizes our findings. It shows the defect evolution as well as the
Figure 4.21: Defect and dopant behavior during the damage anneal.
diffusion of the shallowest spike, which is located within the Si implant damage region, and a deeper spike.

After the Si implantation, the shallow spike, and part of the deeper one, is found to be inactive. In the figure, the inactive defect is termed, after Jones [73], "boron interstitial cluster" (BIC).

During the initial annealing stages (<15 seconds), the BICs quickly dissolve and provide, presumably, the Si-interstitials for the ultrafast diffusion pulse. In the figure this is illustrated by a broadening of the deeper spike. Also, within this time frame, \{113\} defects nucleate and grow in the heavily damaged region, which is darkly shaded. To explain both the small amount of diffusion and the electrical inactivity of the shallow spike, we assume that most boron atoms make a transition from a BIC to a \{113\} defect, which is more stable.

During prolonged annealing, the \{113\} defects dissolve. The larger defects grow at the expense of the smaller ones. During this process, Si-interstitials are injected into the underlying layers, causing enhanced diffusion of the deeper spike.

Finally, as the \{113\} defects dissolve, they release the captured boron atoms. In the remnants of the \{113\}-dissolution the released boron atoms experience some enhanced diffusion: the shallow spike diffuses again.

4.4.7 Perspectives

Transient enhanced diffusion has stimulated intense interest for both technological and scientific reasons. From the point of view of technology, it is hoped to minimize diffusion effects for shallow junctions, and to obtain more realistic computer-aided predictions of dopant profiles during the development of new manufacturing processes. On the scientific side, the challenge is to understand the kinetics of transient diffusion and electrical activation in terms of microscopic processes within the silicon crystal. The recent ability to grow very pure, nearly defect-free, silicon with sharp doping profiles, in combination with short annealing times by using an RTA apparatus and sensitive analysis techniques such as SIMS, SRP and TEM, has further geared up the research in point defect kinetics.

The present results are especially of interest for forthcoming low thermal budget technologies. TED limits straightforward downscaling of junctions depths which are required in, e.g., the CMOS technology to reduce short-channel effects or in the bipolar technology to achieve high cut-off frequencies.

As an example, we show the impact of TED on an NPN which was made by using SiGe technology at DIMES [110]. A 325 Å strained SiGe base (30% ger-
manium), comprising a boron layer, was grown over the entire wafer. Next, a silicon capping layer of 0.2 μm was grown, which was then preamorphized with an arsenic implant, before the actual phosphorus emitter was implanted. Implanted emitters are favorable to allow easy patterning of devices. For optimal device performance, it is important to confine the boron to the SiGe layer.

![Graphical representation of SIMS profiles](image)

Figure 4.22: SIMS profiles of a DIMES SiGe NPN transistor [111]. The SIMS profiles shown are (a) the as-grown and as-implanted profiles and (b) the annealed profiles (700°C) which show considerable TED. As the SiGe layer appears to block transient phosphorus diffusion, the emitter-base junction is located near the Si/SiGe interface. Actually, these devices put TED into use.

To avoid relaxation of the SiGe layer, the process temperatures are not permitted to exceed 700°C. As equilibrium diffusion is negligible under these conditions, the resulting doping profiles in these devices are determined by TED phenomena. Figure 4.22 gives SIMS profiles of a DIMES SiGe NPN transistor. Typical TED features can be observed: tail diffusion in the phosphorus and out-diffusion of
boron. The phosphorus appears to have been “blocked” by the SiGe layer, which results in a “self-alignment” of the emitter-base junction at the Si-SiGe interface. As the fabrication of these devices actually relies on TED, it is obvious that TED needs to be understood and controlled.

Although it is known that boron diffuses slower in strained SiGe layers [29], little is known about the native point defect parameters in such layers. The thinness of such layers poses additional problems in the determination of point defect parameters.

It is expected that marker layer experiments will continue to be a very useful tool to study point defect and extended defect kinetics.

In most publications about marker layer experiments, the SIMS measurement technique has been used, often in combination with TEM. However, a combination of SIMS and SRP also provides a powerful tool for investigating the damage annealing and activation process, especially when little or no diffusion has yet taken place. For instance, there is no direct evidence for dopant migration at room temperature. Deactivation measurements then provide an alternative to monitor the damage effects. The ion-implantation induced deactivation experiments by Larsen et al. [88] are an excellent example in which SRP measurements gain information about point defect diffusivities where SIMS would not.

4.5 Vertical DMOST profiles

In this section, we return to the doping profiles of the DMOST and NPN of the L303 process, which is currently in-line at Philips, see also figures 1.2 and 1.3. In chapter 3 we focused on the determination of the lateral doping profiles; in this chapter, we analyze the vertical profiles by means of SIMS. This subject is addressed here because the treatment of —anomalous— diffusion effects in the previous sections is helpful to understand the diffusion behavior of the profiles.

In the SIMS measurements, we observe enhanced diffusion effects during the early stages of the anneal after the high-dose emitter implantation. Moreover, a sharp kink, at the former amorphous-crystalline (a/c) interface, is observed throughout the anneal. These effects are attributed to implantation damage annealing effects.
Table 4.2: Overview of the anneals carried out. The point of departure is the complete SN anneal of the L303 process (profile d) of which ‘snapshots’ were taken. Figure 4.23 shows schematically the anneal history of the various profiles.

<table>
<thead>
<tr>
<th>profile</th>
<th>SN implant</th>
<th>stabilization</th>
<th>ramp up</th>
<th>anneal</th>
<th>ramp down</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>d</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>e</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f</td>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

4.5.1 Experimental setup

The following processing steps were carried out:

1. starting material: p-type, low resistivity <111> silicon;
2. 70 nm grown oxide (the gate oxide);
3. SP (shallow p) boron implantation: 7×10^{13} \, \text{cm}^{-2}, 40 \, \text{keV};
4. SP anneal: 480 minutes at 1150 °C;
5. SN (shallow n) phosphorus implantation: 5×10^{15} \, \text{cm}^{-2}, 130 \, \text{keV};
6. various furnace anneals according table 4.2.

The point of departure was the full SN anneal of the L303 process. However, to study the profiles during this anneal, the batch was split after the SN implantation and subjected to various anneals which were parts of the full annealing scheme. Figure 4.23 shows schematically the anneals which are summarized in table 4.2. Profiles a through d received the SN implantation, e and f are reference wafers with only boron to study the effect of phosphorus on boron.

4.5.2 Measurement results

Figure 4.24 shows the SIMS measurements from which some remarkable features may be noted.
Implant damage annealing effects in the tail region

First, after the stabilization anneal of 5 minutes at 800°C, the phosphorus profile has a pronounced tail and a small ‘plateau’ at the former a/c interface (at a depth of approximately 0.18 μm), whereas boron shows a pile-up at the same location. In the tail region the diffusion coefficient is found to be \( \approx 3 \times 10^{-13} \text{ cm}^2\text{s}^{-1} \), which is more than four orders of magnitude higher than the intrinsic phosphorus diffusion coefficient at 800°C (\( \approx 1 \times 10^{-17} \text{ cm}^2\text{s}^{-1} \), [39]). This is a clear signature of transient enhanced diffusion beyond the a/c interface, as discussed in section 4.4. The plateau in the phosphorus and the pile-up of boron reflect the gettering capability of the EOR defects. The small ‘plateau’ is also measured for high dose amorphizing boron (BF₂) implants (see e.g. [40, 84]) and is not a typical phosphorus feature.

As the diffusivity is only enhanced beyond the EOR defects, both boron and phosphorus are supplied from depth rather than from the surface region. The driving force for boron might be the well-known electric field effect (see, e.g., Fair [39]), under extrinsic diffusion conditions, which pulls boron towards the EOR defects. Sadana et al. [136], however, found that boron segregation is driven primarily by the implantation-induced damage by comparing amorphizing arsenic and germanium implantations into B-doped silicon. Moreover, the electric field drives phosphorus into the bulk and this effect can therefore not explain the plateau. If we assume that these observations are due to damage annealing effects, we have two explanations.
The first assumes that dislocation loops are effective sinks, not only for silicon interstitials, but also for boron and phosphorus. Uwasawa et al. [169], for instance, developed a boron diffusion model including segregation into EOR defects. To verify their model, they used, like Sadana et al. [136], both high dose arsenic and germanium implants to create EOR defects. They also observed segregation of boron in both cases, thus providing additional evidence that the electric field is not the only contributor to boron pile-up at the a/c interface.

The second explanation assumes that a large interstitial gradient is present at
the EOR location during the point defect cluster dissolution. It is known that these gradients act as a driving force for dopant redistribution [131]. Kim et al. [84] have shown in a modeling exercise for an amorphizing boron (BF$_2$) profile that these gradients provide an alternative explanation for the appearance of a small plateau.

**Phosphorus 'kink' during the anneals**

Second, a sharp 'kink' is observed in all phosphorus profiles, located exactly at the a/c interface. Although high dose phosphorus profiles in defect-free silicon also exhibit a —less sharp— kink (see section 4.3.2) this is attributed to the presence of EOR defects. It is known that EOR dislocation loops dissolve at temperatures higher than 1000 °C [77], and although we did not confirm it directly by means of TEM, we assume their presence in our measurements.

### 4.5.3 A modeling exercise

In an attempt to model the profiles, we used the process simulator TSUPREM4 [166]. The version which we used did not yet have a model for dislocation loop growth and dissolution, nor for dopant gettering. Therefore, an artificial interface was defined at the former a/c interface in order to simulate the EOR defects. This also enabled us to deal with the transient enhanced diffusion beyond the EOR defects.

We used the as-implanted SIMS measurements as a starting profile and scaled the other profiles vertically to match the doses. The implantation damage is included by assuming the +1 model (see section 4.3.3) beyond the a/c interface. The areal density of the +1 interstitials was approximately 3×10$^{14}$ cm$^{-2}$. In the default simulations, no implantation damage was included.

Figure 4.25 shows the default TSUPREM4 simulations together with the measured profiles. Although the junction delineation after the entire anneal is excellent, the kink is not simulated. Introduction of the artificial interface gave the profiles shown in figure 4.26. As can be observed, all characteristic features are simulated.

The TED effects during the stabilization are not modeled properly as this TSUPREM4 version did not have a model for point defect cluster dissolution. The transient in the simulation is caused by free-moving +1 interstitials instead of dissolving defects. However, it shows the phosphorus plateau at the a/c interface, driven by a gradient in the interstitial concentration, as well as TED in the phosphorus tail and a pronounced boron pile-up. The modeling exercise provides a consistent
Figure 4.25: Default TSUPREM4 simulation results. Figure (b) is a detail of the top of the phosphorus profile in (a). For clarity, only two boron profiles are shown.

view of the diffusion behavior of a high dose phosphorus implant in the vertical direction. It is concluded that EOR defects play an important role in determining the profile at high concentrations.
Figure 4.26: Simulations with EOR damage effects, included by defining a sink for interstitials at the EOR location (at 0.18 μm).

4.5.4 Enhanced boron diffusion due to damage annealing after a phosphorus implantation

Although the TED did not strongly affect the final vertical boron profile, it may become different if a large gradient in the boron profile is present. Due to segregation effects, during earlier processing, such a gradient exists in the boron profile near the surface. We here now demonstrate that sharp boron profiles do indeed smear out during a short, low temperature anneal after a phosphorus ion implantation. The experiment was done together with the enhanced diffusion experiments,
which were discussed in section 4.4.

We used structure A (see section 4.4.3 and figure 4.7), with the second spike being the high one, and implanted it with phosphorus. The dose was chosen to be equal to the SN implant dose ($5 \times 10^{15}$ cm$^{-2}$). We could not grow the 70 nm ox-

![Graph showing boron enhanced diffusion during the anneal of a 15 keV, $5 \times 10^{15}$ cm$^{-2}$ phosphorus implant. The annealing time and temperature were 5 minutes and 700°C respectively. This sample was part of batch A, cf. figures 4.7 through 4.9.]

Figure 4.27: Illustration of boron enhanced diffusion during the anneal of a 15 keV, $5 \times 10^{15}$ cm$^{-2}$ phosphorus implant. The annealing time and temperature were 5 minutes and 700°C respectively. This sample was part of batch A, cf. figures 4.7 through 4.9.

de in order to avoid oxidation enhanced diffusion of the spikes and, therefore, we implanted the phosphorus in the bare wafer with reduced energy (15 keV instead of 130 keV). The anneal was done in a furnace, at 700°C, via a quick manual load to avoid the cooling down of the boat which would normally occur during loading and unloading. In the processing of the L303 process, the average temperature during the loading and stabilization period is lower than 800°C, and we therefore think that the results from both experiments can be compared qualitatively.

In figure 4.27, the SIMS measurement of the boron spikes is shown. Obviously, the phosphorus implant resulted in pronounced TED of all boron spikes, and also induced boron clustering. The averaged diffusivity enhancement in the boron spikes ($>3 \times 10^4$) is even higher than in the diffusion experiments given in the previous chapter, which is quite intriguing but has not been analyzed further.

We conclude that phosphorus implants cause pronounced TED in its tail and in—sharp— boron profiles at moderate annealing temperatures. In the DMOST,
the channel region is adjacent to the phosphorus profile and is, therefore, sensitive to implantation damage annealing effects. This might cause a redistribution of the doping profile near the Si-SiO₂ interface, which has an immediate effect on, e.g., the threshold voltage. This could explain, at least in part, why the electrical measurements in section 2.6 could be satisfactorily described with unsegregated doping profiles.
4. Transient enhanced diffusion
Chapter 5

Conclusion

While aiming at the simultaneous optimization of a vertical double diffused MOS (VDMOS) and an NPN transistor with TCAD tools, we encountered several issues which have been addressed in this thesis.

To reduce the large computation times, an efficient optimization strategy has been proposed. Next, to characterize the doping profiles along the interface in the VDMOS, two measurement methods have been developed, one direct and quick, and one based on inverse modeling. Finally, transient enhanced boron diffusion experiments were carried out in order to better understand implantation damage annealing effects.

In the following section, we list the conclusions by headers of the three subjects. Suggestions for further research are given in the final section.

5.1 Conclusions

Optimization strategy

A new optimization strategy to extract maximum information from a minimal set of numerical experiments while arriving at a global optimum has been proposed. It combines the strength of two existing methods, i.e., the Response Surface Model and Levenberg-Marquardt method, where the former is used to collect \textit{a priori} information for an efficient fine-tuning by the latter. Two examples have been provided to illustrate the strategy:

- As a fast and simple example, the calibration of a compact MOS model on the linear characteristics of a lateral DMOS transistor is shown;
The feasibility of complex device optimizations is illustrated by a simultaneous optimization of the VDMOS and NPN transistor.

**Determination of doping profiles**

Two nondestructive, electrical, methods have been proposed to measure lateral doping profiles.

The first is based on a threshold voltage formulation ("$V_T$ method") and is essentially a 1D profiling method, applicable to a gated diode structure. As it requires only standard $CV$ measurement equipment and no complicated calculations, the method allows a fast inspection of lateral doping profiles such as the channel of the VDMOST or the source and drain profiles of a MOSFET. The following conclusions apply:

- Two test devices have been designed, processed and measured. Both devices had a lateral DMOST structure, one complete and one without the source, the latter being essentially a gated diode. The measurements show excellent sensitivity to dose variations of the channel profile;

- The method is limited to a maximum doping level of approximately $1 \times 10^{19}$ cm$^{-3}$, because at higher doping concentrations oxide breakdown occurs before an inversion layer is formed;

- The 1D assumption results in an overestimation near the lateral junction, and a sharp pile-up in the profile near the gate edge. The latter allows the positioning of the profile relative to the gate edge;

- A charge balance analysis near the gate edge yields the rule of thumb that the oxide thickness must considerably exceed the lateral out-diffusion length to justify the 1D assumption.

The second method is based on the inverse modeling approach and does not require the doping profiles to be expressed in the measured variable, unlike the first method. During inverse modeling, a parameterized doping profile is iteratively improved by an optimizer by minimizing the difference between device simulation and experiment. Spline functions are used as a parameterized doping profile model because they can take arbitrary, continuous, shapes with a limited number of parameters. This allows inverse modeling to be still feasible without requiring excessive computation times. The following conclusions apply:
5.1. Conclusions

- Although the method is not essentially 1D, we only extracted a net lateral doping profile beneath the gate of the gated diode. The DMOS test devices were originally designed to test the $V_T$ method and were not suited to collect enough electrical information for a 2D characterization;

- During the extraction procedure, the optimizer repeatedly invokes a dedicated spline program to calculate doping profile updates. This program filters a "TIF" (Technology Interface Format) file, which describes the complete two-dimensional device geometry, thereby replacing the doping concentrations by their updated values. The altered TIF file is then supplied to the device simulator. The communication through TIF files provides the advantage that complex device geometries can be used, as, e.g., calculated by a TIF-supporting process simulator such as TSUPREM4.

Transient enhanced diffusion experiments

The following conclusions are drawn from the experiments on transient enhanced diffusion of $\delta$-doped boron layers during annealing, at 700 °C, of damage which was created by nonamorphizing silicon implantations.

- A very fast diffusion transient is observed. The diffusion enhancement during the first 15 seconds of the anneal is approximately $1 \times 10^5$, which is a time-averaged value and thus a lower limit;

- The short annealing time of 15 seconds gives a lower limit for the diffusivity of the silicon self-interstitial ($2 \times 10^{-10}$ cm$^2$s$^{-1}$) and boron interstitial ($2 \times 10^{-13}$ cm$^2$s$^{-1}$);

- Spreading resistance profiling measurements showed boron deactivation up to depths of 0.5 μm, induced by the silicon self-implantation. The absence of boron migration indicates that the inactive boron resides in defects which contain only one boron atom;

- During annealing, the profiles reactivated within 15 seconds. This indicates that the inactive boron atoms resided in relatively unstable defects. The boron atoms in the implant damage region were, however, driven into immobile clusters again. These clusters were more persistent and appeared to anneal out on the same time scale as the implant damage.
We also looked at the diffusion behavior of the phosphorus and boron profiles in the DMOS process. A limited set of measurements and the complicated kinetics associated with amorphizing phosphorus implantations do not permit such precise analysis as could be done in the specially designed test structures of the enhanced diffusion experiments. The following conclusions can, however, be drawn, some directly, and some after comparison with literature:

- During annealing, end-of-range defects were formed; they plaid an important role in determining the phosphorus profile at high concentrations;

- During the loading phase of the anneal (5 minutes at 800 °C), enhanced phosphorus diffusion was found only beyond the location of the EOR-defects, which is consistent with earlier-reported observations.

- The EOR defects getter boron and phosphorus during the loading phase, which resulted in boron pile-up and a "plateau" in the phosphorus profile. The dopants were supplied from depth rather than from the surface region, which is also consistent with the concept of TED beyond the EOR defects.

- The impact of TED on the vertical boron profile was not significant, as its concentration gradient is small. We speculate, however, that the channel region of the DMOS is sensitive to TED, as a boron gradient may exist as a result of segregation during earlier processing. Indirect proof for was given by one of the boron marker layer experiments in which one wafer was implanted with phosphorus instead of silicon. The sharp marker layers have diffused considerably during an anneal at 700 °C for 5 minutes.

5.2 Recommendations for future work

Suggestions for future work:

- Further development of the spline profiling method is needed. The method still requires considerable effort mainly because of the critical knot placement, which is awkward to include in the extraction procedure. We have used the Levenberg-Marquardt optimization method to determine the spline coefficients. However, without sufficient a priori knowledge, the optimization strategy proposed in chapter 2 may be applied.

- Diffusion experiments:
5.2. Recommendations for future work

- In view of the room temperature mobility of the silicon self-interstitial, more experiments, at relatively low temperatures, are required to determine the point defect parameters;
- Measurements with boron-doped and intrinsic capping layers are necessary to determine the role of boron in the ultrafast transient;
- Adequate damage annealing models are required for accurate process simulation tools;
- Enhanced diffusion phenomena in strained Si/SiGe heterostructures are still not well understood. The anticipated importance of SiGe technology calls for more and dedicated experiments.
5. Conclusion
References


References


References


[75] K.S. Jones, J. Liu, L. Zhang, V. Krishnamoorthy, and R.T. DeHoff. Studies of the interactions between (311) defects and type I and II dislocation loops


List of symbols and abbreviations

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_I$</td>
<td>dopant interstitialcy</td>
<td></td>
</tr>
<tr>
<td>$A_i$</td>
<td>interstitial dopant atom</td>
<td></td>
</tr>
<tr>
<td>$A_s$</td>
<td>substitutional dopant atom</td>
<td></td>
</tr>
<tr>
<td>$AV$</td>
<td>dopant vacancy pair</td>
<td></td>
</tr>
<tr>
<td>$a$</td>
<td>boundary knot in knot sequence</td>
<td>cm</td>
</tr>
<tr>
<td>$a_{ij}$</td>
<td>response surface model term</td>
<td>cm</td>
</tr>
<tr>
<td>$a_T$</td>
<td>trapping radius</td>
<td>cm</td>
</tr>
<tr>
<td>$C$</td>
<td>capacitance</td>
<td>F</td>
</tr>
<tr>
<td>$B_{clus}$</td>
<td>boron-interstitial cluster</td>
<td></td>
</tr>
<tr>
<td>$B_i^c$</td>
<td>interstitial boron atom at charge state $c$</td>
<td></td>
</tr>
<tr>
<td>$B_s^+$</td>
<td>substitutional boron atom</td>
<td></td>
</tr>
<tr>
<td>$BV_{DS}$</td>
<td>drain-source breakdown voltage</td>
<td>V</td>
</tr>
<tr>
<td>$b$</td>
<td>boundary knot in knot sequence</td>
<td>cm</td>
</tr>
<tr>
<td>$C_{GB}$</td>
<td>gate bulk capacitance</td>
<td>F/cm</td>
</tr>
<tr>
<td>$C_{GD}$</td>
<td>gate drain capacitance</td>
<td>F/cm</td>
</tr>
<tr>
<td>$C_{GS}$</td>
<td>gate source capacitance</td>
<td>F/cm</td>
</tr>
<tr>
<td>$C_{ox}$</td>
<td>oxide capacitance</td>
<td>F/cm$^2$</td>
</tr>
<tr>
<td>$C_{le}$</td>
<td>lower edge fringing capacitance</td>
<td>F/cm</td>
</tr>
<tr>
<td>$C_{sw}$</td>
<td>side wall fringing capacitance</td>
<td>F/cm</td>
</tr>
<tr>
<td>$C_{top}$</td>
<td>top capacitance</td>
<td>F/cm</td>
</tr>
<tr>
<td>$C_X^c$</td>
<td>concentration of point defect $X$</td>
<td>cm$^{-3}$</td>
</tr>
<tr>
<td>$c$</td>
<td>charge state of point defect $X$</td>
<td></td>
</tr>
<tr>
<td>$c_i$</td>
<td>$i$-th $B$-spline coefficient in spline function</td>
<td></td>
</tr>
<tr>
<td>$D_X$</td>
<td>diffusivity of point defect $X$</td>
<td>cm$^2$/s</td>
</tr>
<tr>
<td>$f_T$</td>
<td>cut-off frequency</td>
<td>Hz</td>
</tr>
<tr>
<td>Symbol</td>
<td>Description</td>
<td>Unit</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td>------</td>
</tr>
<tr>
<td>$g$</td>
<td>number of interior knots in knot sequence</td>
<td></td>
</tr>
<tr>
<td>$h_{FE}$</td>
<td>dc current gain</td>
<td></td>
</tr>
<tr>
<td>$I$</td>
<td>silicon self-interstitial</td>
<td></td>
</tr>
<tr>
<td>$I^c$</td>
<td>silicon self-interstitial at charge state $c$</td>
<td></td>
</tr>
<tr>
<td>$I$</td>
<td>Electrical current</td>
<td>A</td>
</tr>
<tr>
<td>$K$</td>
<td>bimolecular recombination rate of silicon interstitials and vacancies</td>
<td>s$^{-1}$</td>
</tr>
<tr>
<td>$k$</td>
<td>Boltzmann’s constant ($1.38 \times 10^{-23}$)</td>
<td>J/K</td>
</tr>
<tr>
<td>$k$</td>
<td>degree of B-spline</td>
<td></td>
</tr>
<tr>
<td>$L$</td>
<td>extracted lateral doping axis</td>
<td>cm</td>
</tr>
<tr>
<td>$L'$</td>
<td>average characteristic delay length</td>
<td>cm</td>
</tr>
<tr>
<td>$L_G$</td>
<td>length of noninverted part of the gate</td>
<td>cm</td>
</tr>
<tr>
<td>$L_{OV}$</td>
<td>drawn gate length</td>
<td>cm</td>
</tr>
<tr>
<td>$N$</td>
<td>diffusion length under gate</td>
<td>cm</td>
</tr>
<tr>
<td>$(g$-gate-drain or gate-source overlap)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N_A$</td>
<td>acceptor concentration</td>
<td>cm$^{-3}$</td>
</tr>
<tr>
<td>$N_D$</td>
<td>donor concentration</td>
<td>cm$^{-3}$</td>
</tr>
<tr>
<td>$N_{epi}$</td>
<td>epi-layer concentration</td>
<td>cm$^{-3}$</td>
</tr>
<tr>
<td>$N_i$</td>
<td>$i$-th B-spline</td>
<td></td>
</tr>
<tr>
<td>$N_N$</td>
<td>net doping concentration</td>
<td>cm$^{-3}$</td>
</tr>
<tr>
<td>$N_{Ptail}^{surf}$</td>
<td>surface concentration of phosphorus tail</td>
<td>cm$^{-3}$</td>
</tr>
<tr>
<td>$N_T$</td>
<td>interstitial trap concentration</td>
<td>cm$^{-3}$</td>
</tr>
<tr>
<td>$n_i$</td>
<td>intrinsic carrier concentration</td>
<td>cm$^{-3}$</td>
</tr>
<tr>
<td>$n_x$</td>
<td>$x$-th $n$-type doping spline coefficient</td>
<td>cm$^{-3}$</td>
</tr>
<tr>
<td>$p$</td>
<td>parameter vector</td>
<td></td>
</tr>
<tr>
<td>$p_x$</td>
<td>$x$-th $p$-type doping spline coefficient</td>
<td>cm$^{-3}$</td>
</tr>
<tr>
<td>$Q_B$</td>
<td>bulk charge</td>
<td>C</td>
</tr>
<tr>
<td>$q$</td>
<td>magnitude of electronic charge ($1.6 \times 10^{-19}$)</td>
<td>C</td>
</tr>
<tr>
<td>$R$</td>
<td>response</td>
<td></td>
</tr>
<tr>
<td>$R_{acc}$</td>
<td>accumulation layer resistance</td>
<td>$\Omega$</td>
</tr>
<tr>
<td>$R_b$</td>
<td>bulk resistance</td>
<td>$\Omega$</td>
</tr>
<tr>
<td>$R_{sp}$</td>
<td>specific on-resistance</td>
<td>$\Omega$mm$^2$</td>
</tr>
<tr>
<td>$R_{on}$</td>
<td>constant series drain-source resistance</td>
<td>$\Omega$</td>
</tr>
<tr>
<td>$R_{\square}$</td>
<td>sheet resistance</td>
<td>$\Omega$/$\square$</td>
</tr>
<tr>
<td>Symbol</td>
<td>Description</td>
<td>Unit</td>
</tr>
<tr>
<td>--------</td>
<td>--------------------------------------------------</td>
<td>----------</td>
</tr>
<tr>
<td>$s$</td>
<td>spline function</td>
<td></td>
</tr>
<tr>
<td>$t_{ox}$</td>
<td>gate oxide thickness</td>
<td>cm</td>
</tr>
<tr>
<td>$T$</td>
<td>temperature</td>
<td>K</td>
</tr>
<tr>
<td>$T_{ox}$</td>
<td>gate oxidation time</td>
<td>s</td>
</tr>
<tr>
<td>$T_{SP}$</td>
<td>shallow P drive in time</td>
<td>s</td>
</tr>
<tr>
<td>$t_{epi}$</td>
<td>epi-layer thickness</td>
<td>cm</td>
</tr>
<tr>
<td>$U_A$</td>
<td>Early voltage</td>
<td>V</td>
</tr>
<tr>
<td>$V$</td>
<td>voltage</td>
<td>V</td>
</tr>
<tr>
<td>$V$</td>
<td>vacancy</td>
<td></td>
</tr>
<tr>
<td>$V_{FB}$</td>
<td>flatband voltage</td>
<td>V</td>
</tr>
<tr>
<td>$V_T$</td>
<td>threshold voltage</td>
<td>V</td>
</tr>
<tr>
<td>$V_{t0}$</td>
<td>zero backbias threshold voltage</td>
<td>V</td>
</tr>
<tr>
<td>$W$</td>
<td>device width</td>
<td>cm</td>
</tr>
<tr>
<td>$\vec{w}$</td>
<td>vector with weights</td>
<td></td>
</tr>
<tr>
<td>$\vec{y}_{meas}$</td>
<td>measured data vector</td>
<td></td>
</tr>
<tr>
<td>$\vec{y}_{calc}$</td>
<td>calculated data vector</td>
<td></td>
</tr>
<tr>
<td>$\beta$</td>
<td>gain factor</td>
<td>A/V²</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>dielectric permittivity</td>
<td>F/cm</td>
</tr>
<tr>
<td>$\Gamma$</td>
<td>body effect coefficient</td>
<td>V¹/²</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>(surface) body effect coefficient</td>
<td>V¹/²</td>
</tr>
<tr>
<td>$\gamma_i$</td>
<td>bulk body effect coefficient</td>
<td></td>
</tr>
<tr>
<td>$\lambda$</td>
<td>Marquardt parameter</td>
<td>1</td>
</tr>
<tr>
<td>$\lambda_i$</td>
<td>position of $i$-th knot in knot sequence</td>
<td>cm</td>
</tr>
<tr>
<td>$\mu$</td>
<td>carrier mobility</td>
<td>cm²/Vs</td>
</tr>
<tr>
<td>$\Phi_{SN}$</td>
<td>shallow N implant dose</td>
<td>cm⁻²</td>
</tr>
<tr>
<td>$\Phi_{SP}$</td>
<td>shallow P implant dose</td>
<td>cm⁻²</td>
</tr>
<tr>
<td>$\phi_F$</td>
<td>bulk potential</td>
<td>V</td>
</tr>
<tr>
<td>$\sigma_x$</td>
<td>vertical characteristic length of a doping profile</td>
<td>cm</td>
</tr>
<tr>
<td>$\sigma_y$</td>
<td>horizontal characteristic length of a doping profile</td>
<td>cm</td>
</tr>
<tr>
<td>$\theta_A$</td>
<td>mobility reduction factor due to a vertical electric field</td>
<td>V⁻¹</td>
</tr>
<tr>
<td>$\theta_B$</td>
<td>mobility reduction factor due to a lateral electric field</td>
<td>V⁻¹/²</td>
</tr>
<tr>
<td>$\xi$</td>
<td>Least squares error</td>
<td></td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Meaning</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>a/c interface</td>
<td>amorphous-crystalline interface</td>
<td></td>
</tr>
<tr>
<td>BIC</td>
<td>boron interstitial cluster</td>
<td></td>
</tr>
<tr>
<td>CV</td>
<td>capacitance voltage</td>
<td></td>
</tr>
<tr>
<td>CVD</td>
<td>chemical vapor deposition</td>
<td></td>
</tr>
<tr>
<td>DMOS(T)</td>
<td>double diffused MOS (transistor)</td>
<td></td>
</tr>
<tr>
<td>DOE</td>
<td>design of experiments</td>
<td></td>
</tr>
<tr>
<td>EOR</td>
<td>end-of-range</td>
<td></td>
</tr>
<tr>
<td>IM</td>
<td>inverse modeling</td>
<td></td>
</tr>
<tr>
<td>LDMOST</td>
<td>lateral double diffused MOS transistor</td>
<td></td>
</tr>
<tr>
<td>LM</td>
<td>Levenberg-Marquardt</td>
<td></td>
</tr>
<tr>
<td>MOS</td>
<td>'metal'-oxide-semiconductor</td>
<td></td>
</tr>
<tr>
<td>OED</td>
<td>oxidation enhanced diffusion</td>
<td></td>
</tr>
<tr>
<td>RSM</td>
<td>response surface model</td>
<td></td>
</tr>
<tr>
<td>RTA</td>
<td>rapid thermal annealing</td>
<td></td>
</tr>
<tr>
<td>RPCVD</td>
<td>reduced pressure chemical vapor deposition</td>
<td></td>
</tr>
<tr>
<td>SIMS</td>
<td>secondary ion mass spectrometry</td>
<td></td>
</tr>
<tr>
<td>SRP</td>
<td>spreading resistance profiling</td>
<td></td>
</tr>
<tr>
<td>STM</td>
<td>scanning tunneling microscope</td>
<td></td>
</tr>
<tr>
<td>TCAD</td>
<td>technology computer aided design</td>
<td></td>
</tr>
<tr>
<td>TED</td>
<td>transient enhanced diffusion</td>
<td></td>
</tr>
<tr>
<td>TEM</td>
<td>transmission electron microscopy</td>
<td></td>
</tr>
<tr>
<td>TIF</td>
<td>technology interface format</td>
<td></td>
</tr>
<tr>
<td>TMA</td>
<td>Technology Modeling Associates inc.</td>
<td></td>
</tr>
<tr>
<td>VDMOST</td>
<td>vertical double diffused MOS transistor</td>
<td></td>
</tr>
</tbody>
</table>
Summary

In this thesis the simultaneous optimization of a Vertical Double Diffused MOS (VDMOS) and an NPN transistor is investigated by using technology computer aided design tools. Both devices are fabricated with the same processing steps which complicates the device optimization. The problem requires a two-dimensional approach as the VDMOST is defined by laterally out-diffusing doping profiles.

Because of the uncertainties in process modeling of 2D dopant profiles and large computation times during optimization, we, in this thesis, not only present an efficient optimization strategy, but also new doping profile measurement methods. We also present new experiments investigating enhanced diffusion of boron during ion implantation damage annealing.

Despite increasing computer power, simulations for process optimization still require appreciable times making typical 2D optimizations easily become impractical. We have therefore merged two existing optimization methods, i.e. the Response Surface Model (RSM) method and the Levenberg-Marquardt (LM) method. This gives efficient optimizations yielding accurate optima with a limited set of numerical experiments.

The global approach of the RSM-method is used to screen out less sensitive parameters and to generate initial solutions. The RSM-method however, relies on approximating analytical models which do not necessarily provide sufficient accuracy. Rather than trying to improve these models, we propose to subsequently switch to the gradient-based LM-method which does not rely on approximating models. In this way, the RSM-method is essentially used to collect a priori information for an efficient LM-iteration.

The accuracy of the optimization depends on the accuracy of the simulated 2D doping profiles which, in general, is questionable. Because no general 2D profile measurement methods are available process simulators are not well-calibrated in
two dimensions. In order to characterize the doping profiles in the VDMOST, we present two nondestructive, electrical, measurement methods.

The first method is based on a threshold voltage formulation and uses \( CV \)-measurements only ("\( V_T \) method"). It is essentially a 1D formulation and it allows a fast inspection of the lateral doping profiles along the interface using standard \( CV \)-measurement equipment. This method is very sensitive and is therefore useful during process development and for in-line use.

The second method is based on the more general inverse modeling approach in which nonlinear optimization techniques are used to iteratively improve a parameterized doping profile by matching simulated with measured electrical device characteristics. Inverse modeling is more complex than the first method but does not rely on simplifying assumptions and is not restricted to one spatial dimension. A general drawback however, is that unmodeled features in the doping profiles cannot be resolved and its applicability is therefore restricted to well understood and modeled phenomena. To overcome this problem we used spline functions to describe the doping profile. Spline functions take arbitrary, continuous, shapes with a limited number of parameters, allowing inverse modeling to be still feasible without requiring excessive computation times.

In the course of the numerical optimizations, it was found that the VDMOST doping profiles showed features attributed to the annealing of implantation damage, such as transient enhanced diffusion. Although of less importance for the final vertical VDMOST profiles, enhanced diffusion phenomena have become a key issue in the activation of implanted profiles in state of the art IC-processes. These effects are not well understood and modeled. Therefore, implantation damage annealing experiments have been initiated which were further motivated by the ability to experiment with very pure silicon epitaxial layers, incorporating narrow boron-doped layers, using the ASM-Epsilon Chemical Vapour Deposition reactor.

Transient enhanced diffusion is associated with damage annealing which is known to inject silicon self-interstitials into the substrate thereby establishing a temporary supersaturation of silicon interstitials. As boron is known to diffuse primarily by interaction with silicon interstitials, its diffusion is enhanced during the damage annealing. In our experiments, the damage annealing process is therefore monitored by measuring the diffusive behaviour of boron marker layers. The implant damage is created by silicon self-implantations as the amount of excess interstitials to be considered is then relatively well known from literature.

The results reveal a short, but fierce, diffusion transient which precedes earlier
reported transient enhanced diffusion regimes. The short annealing times allow the calculation of a lower limit, at 700 °C, for the much disputed silicon interstitial diffusivity. Moreover, boron deactivation due to the silicon implant and during the damage annealing process is observed and analyzed.
Samenvatting

In dit proefschrift is de gelijktijdige optimalisatie van een verticale dubbel-gediffundeerde “power MOS” (VDMOS) en NPN transistor bestudeerd. Beide componenten zijn gefabriceerd met dezelfde processtappen hetgeen de optimalisatie ervan compliceert. Het probleem eist een twee-dimensionale aanpak omdat de linearbreak VDMOS transistor gemaakt wordt door lateraal uitdiffunderende doping profielen.

Vanwege de onnauwkeurigheid in de modellerings van twee-dimensionale doping profielen en grote rekentijden tijdens de optimalisatie, presenteren we in dit proefschrift niet alleen een efficiëntere optimalisatie strategie maar ook een nieuwe methoden om het doping profiel te meten. Bovendien presenteren we nieuwe experimenten om versnelde diffusie van boor tijdens het uitstoken van schade na een ionen implantatie te bepalen.

Ondanks de toenemende rekenkracht van computers, verbruiken simulatoren nog steeds aanzienlijk rekentijden en typische twee-dimensionale optimalisaties worden al snel onhandelbaar. Om deze reden hebben we twee optimalisatie methoden samengevoegd, te weten de Response Surface Model (RSM) - en Levenberg-Marquardt (LM) methoden. Hiermee zijn efficiënte optimalisaties mogelijk die nauwkeurige minima opleveren met een beperkte hoeveelheid numerieke experimenten.

De globale benadering van de RSM-methode wordt gebruikt om de gevoelige parameters te bepalen en om initiële oplossingen te genereren. Echter, de RSM-methode vertrouwt op benaderende analytische modellen die niet noodzakelijkerwijs voldoende nauwkeurig zijn. In plaats van deze modellen te verbeteren, stellen we voor om vervolgens over te stappen op de LM-methode, die gebruik maakt van gradienten en niet vertrouwt op benaderende modellen. Op deze wijze wordt de RSM-methode in essentie gebruikt om a priori informatie te verzamelen voor een efficiënte LM-iteratie.
De nauwkeurigheid van de optimalisatie hangt af van de nauwkeurigheid van de gesimuleerde 2D profielen, welke, in het algemeen, twijfelachtig is. Omdat er geen algemene 2D doping profiel meetmethoden beschikbaar zijn, zijn proces simula- toren niet goed gecalibreerd in twee dimensies. Om de doping profielen in de VDMOS transistor te karakteriseren, presenteren we twee niet-destructieve, elektrische, meetmethoden.

De eerste methode is gebaseerd op een drempelspanning-formulering en maakt uitsluitend gebruik van $CV$-metingen ("$V_T$-methode"). Het is een 1D-formulering en staat een snelle inspecte toe van van laterale doping profielen, waarbij gebruik gemaakt wordt van standaard $CV$-meetapparatuur. De methode is erg gevoelig en daarom bruikbaar tijdens procesontwikkeling of voor "in-line"-gebruik.

De tweede methode is gebaseerd op de meer algemene "inverse modeling" be- nadering. Hierbij worden niet-lineaire optimalisatie technieken gebruikt om iteratief een geparameerseerd doping profiel te verbeteren door het verschil tussen metingen en simulaties te minimaliseren. Deze methode is complexer dan de eerre, maar vertuigt niet op vereenvoudigde benaderingen en is niet beperkt tot één ruimtelijke dimensie. Een algemene tekortkoming echter is dat niet-gemodelleerde kenmerken in de doping profielen niet zichtbaar kunnen worden gemaakt, en de toepasbaarheid is daarom beperkt tot goed begrepen en gemodelleerde fenome- nen. Om dit probleem te op te lossen, hebben we spline functies gebruikt om de doping profielen te beschrijven. Spline functies kunnen elke willekeurige, conti- nue, vorm aannemen met een beperkte hoeveelheid parameters, waardoor inverse modellering nog steeds haalbaar is zonder dat excessieve rekentijden nodig zijn.


"Transient enhanced diffusion" wordt geassocieerd met het uitstoken van scha-
De waarvan bekend is dat silicium zelf-interstitiëlen worden geïnjecteerd in het substraat. Dit veroorzaakt een tijdelijke oververzadiging van silicium interstitiëlen. Omdat bekend is dat boor voornamelijk diffundeert door interactie met silicium interstitiëlen, wordt de boor-diffusie verhoogd tijdens het uitstoken van de schade. In onze experimenten wordt daarom het schade uitstook proces geregistreerd door het diffusie-gedrag van meerdere boorlagen te meten. De implantatieschade wordt veroorzaakt door silicium zelf-implantaties omdat de hoeveelheid overtallige interstitiëlen dat dan beschouwd moet worden relatief goed bekend is in de literatuur.

De resultaten laten een korte maar sterke diffusie transient zien die vooraf gaat aan eerder gerapporteerde TED-regimes. Door de korte uitstooktijden is het mogelijk een ondergrens te bepalen De korte annealtijden laten een berekening toe van de ondergrens bij voor de veel betwiste diffusiecoëfficiënt van het silicium interstitiëel, bij 700 °C. Bovendien is deactivatie van boor geobserveerd, zowel als gevolg van de silicium implantatie, als tijdens de schade uitstook.
Biography

Hendrik Gezienus Albert Huizing was born in Emmen, The Netherlands, on December 23, 1967. After graduating in 1986 from the Atheneum-B of the Gemeentelijke Scholen Gemeeschap in Emmen, he began his studies at the Faculty of Electrical Engineering of the Delft University of Technology. He did his Master's work on inverse modeling with process and device simulator TITAN6 at Centre National d'Etudes des Telecommunications (CNET) in Grenoble, France, after which he graduated with an "ingenieurs diploma" early 1992 (M.Sc.). During his studies, he assisted at introductory practical courses for undergraduate students as a "student assistent". Before starting as a research assistant at DIMES, he set up a practical course on IC technology for undergraduate students at the same laboratory, at which he also instructed for three years. His research was on the characterization and modeling of doping profiles. Mid 1996 he joined Philips Research Laboratories, Eindhoven, The Netherlands, as a Research Scientist.
Acknowledgements

The work in this thesis was performed at the Electrical Component, Technology and Materials laboratory, Faculty of Electrical Engineering, Delft University of Technology, and was financed by the Foundation for Technical Sciences (STW) under project number DEL33.3055. The thesis could not have been finished without the help of many people. In particular I wish to thank:

- Prof. dr. C.I.M. Beenakker for “running” a well-organized laboratory and thereby providing a pleasant and inspiring working environment;

- Dr. ir. W. Crans for providing me with the opportunity to start this work. His contagious enthusiasm and stimulating discussions on “inverse modeling” shaped my scientific interests;

- Prof. dr. ir. J.W. Slotboom for helpful discussions and introduction to many researchers;

- Dr. H. van der Vlist, ing. A.L. Klarenberg, J.H.H.A. Egbers, ir. M.J. Swanenberg and G. van Lijnschoten of Philips Components Nijmegen, for providing test devices and measurement data;

- Dr. M. van Dort, prof. dr. N.E.B. Cowern, dr. P. Stolk, dipl. phys. K.K. Larsen, ir. Ludikhuize and lic. R. Vanoppen of Philips Research Laboratories, Eindhoven, for many forms of support, both theoretical and practical. I particularly wish to thank dr. P. Stolk and prof. dr. N. Cowern for critically reading the chapter on transient enhanced diffusion;

- Prof. dr. P. Zalm, dr. J. van Berkum and ing. R. de Kruijff of Philips Research Laboratories, Eindhoven for carrying out excellent SIMS analyses;
• Ir. L.H.M. Dupas and dr. R. Cartuyvels of IMEC, Leuven, Belgium, for an extensive introduction to response surface modeling and pleasant collaboration;

• Dr. V. Privitera of CNR–IMETEM, Catania, Italy, for SRP measurements;

• The scientific staff of the DIMES-ECTM laboratory, in particular dr. L.K. Nanver for useful discussions on technology and dr. C.C.G. Visser for growing the boron superlattices;

• J.C. Staalenburg. I am very thankful for his prompt technical support on both electrical measurements and computer hardware;

• The DIMES support staff for many forms of practical assistance. I particular value the contribution of ing. S. Bootsma, ing. C. de Boer and R. Mallee for carefully carrying out the enhanced diffusion experiments;

• Dr. T. Smedes and ing. J. Satter for helpful discussions on compact MOS models;

• Colleague Ph.D. students, in particular Joost Willemen and Raymond Hue-ting for many stimulating discussions on physics and more worldly subjects;

• The Master’s degree students, Atbin Sadeghzadeh and Fariba Najipour for their contribution on simulations in this project;

• Mrs. Zaat-Jones for correcting the manuscript of this thesis.

Finally, writing a thesis requires much patience and support from friends and family. I therefore wish to thank the three most important people in my life who have contributed by other than scientifical means, however, not less valued:

• My parents for their many ways of giving support throughout the ‘Delft period’;

• My wife, Ageeth, for her constant support and resigned share in the thesis blues.