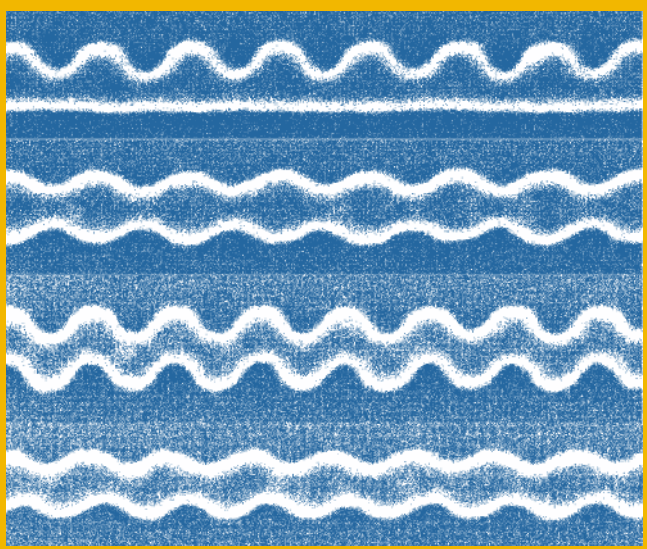
MATCHING PROPERTIES OF DEEP SUB-MICRON MOS TRANSISTORS

Jeroen A. Croon, Willy Sansen
and Herman E. Maes



Springer

MATCHING PROPERTIES OF DEEP SUB-MICRON MOS TRANSISTORS

THE KLUWER INTERNATIONAL SERIES IN ENGINEERING AND COMPUTER SCIENCE

ANALOG CIRCUITS AND SIGNAL PROCESSING

Consulting Editor*: Mohammed Ismail.** ***Ohio State University

Related Titles:

LNA-ESD CO-DESIGN FOR FULLY INTEGRATED CMOS WIRELESS RECEIVERS
Leroux and Steyaert
Vol. 843, ISBN: 1-4020-3190-4

SYSTEMATIC MODELING AND ANALYSIS OF TELECOM FRONTENDS AND THEIR BUILDING BLOCKS
Vanassche, Gielen, Sansen
Vol. 842, ISBN: 1-4020-3173-4

LOW-POWER DEEP SUB-MICRON CMOS LOGIC SUB-THRESHOLD CURRENT REDUCTION
van der Meer, van Staveren, van Roermund
Vol. 841, ISBN: 1-4020-2848-2

WIDEBAND LOW NOISE AMPLIFIERS EXPLOITING THERMAL NOISE CANCELLATION
Bruccoleri, Klumperink, Nauta
Vol. 840, ISBN: 1-4020-3187-4

SYSTEMATIC DESIGN OF SIGMA-DELTA ANALOG-TO-DIGITAL CONVERTERS
Bajdechi and Huijsing
Vol. 768, ISBN: 1-4020-7945-1

OPERATIONAL AMPLIFIER SPEED AND ACCURACY IMPROVEMENT
Ivanov and Filanovsky
Vol. 763, ISBN: 1-4020-7772-6

STATIC AND DYNAMIC PERFORMANCE LIMITATIONS FOR HIGH SPEED D/A CONVERTERS
van den Bosch, Steyaert and Sansen
Vol. 761, ISBN: 1-4020-7761-0

DESIGN AND ANALYSIS OF HIGH EFFICIENCY LINE DRIVERS FOR Xdsl
Piessens and Steyaert
Vol. 759, ISBN: 1-4020-7727-0

LOW POWER ANALOG CMOS FOR CARDIAC PACEMAKERS
Silveira and Flandre
Vol. 758, ISBN: 1-4020-7719-X

MIXED-SIGNAL LAYOUT GENERATION CONCEPTS
Lin, van Roermund, Leenaerts
Vol. 751, ISBN: 1-4020-7598-7

HIGH-FREQUENCY OSCILLATOR DESIGN FOR INTEGRATED TRANSCEIVERS
Van der Tang, Kasperkovitz and van Roermund
Vol. 748, ISBN: 1-4020-7564-2

CMOS INTEGRATION OF ANALOG CIRCUITS FOR HIGH DATA RATE TRANSMITTERS
DeRanter and Steyaert
Vol. 747, ISBN: 1-4020-7545-6

SYSTEMATIC DESIGN OF ANALOG IP BLOCKS
Vandenbussche and Gielen
Vol. 738, ISBN: 1-4020-7471-9

SYSTEMATIC DESIGN OF ANALOG IP BLOCKS
Cheung and Luong
Vol. 737, ISBN: 1-4020-7466-2

LOW-VOLTAGE CMOS LOG COMPANDING ANALOG DESIGN
Serra-Graells, Rueda and Huertas
Vol. 733, ISBN: 1-4020-7445-X

CIRCUIT DESIGN FOR WIRELESS COMMUNICATIONS
Pun, Franca and Leme
Vol. 728, ISBN: 1-4020-7415-8

DESIGN OF LOW-PHASE CMOS FRACTIONAL-N SYNTHESIZERS
DeMuer and Steyaert
Vol. 724, ISBN: 1-4020-7387-9

MODULAR LOW-POWER, HIGH SPEED CMOS ANALOG-TO-DIGITAL CONVERTER FOR EMBEDDED SYSTEMS
Lin, Kemna and Hosticka
Vol. 722, ISBN: 1-4020-7380-1

MATCHING PROPERTIES OF DEEP SUB-MICRON MOS TRANSISTORS

by

Jeroen A. Croon

IMEC, Leuven, Belgium

Willy Sansen

Katholieke Universiteit Leuven,
Leuven, Belgium

and

Herman E. Maes

IMEC and Katholieke Universiteit Leuven,
Leuven, Belgium

Springer

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN 0-387-24314-3 (HB)
ISBN 0-387-24313-5 (e-book)

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

Sold and distributed in North, Central and South America
by Springer,
101 Philip Drive, Norwell, MA 02061, U.S.A.

In all other countries, sold and distributed
by Springer,
P.O. Box 322, 3300 AH Dordrecht, The Netherlands.

Printed on acid-free paper

Printed in the Netherlands.

Contents

Preface ix
Acknowledgments xi

1. INTRODUCTION 1
   1.1 Matching analysis 2
   1.2 Importance for circuit design 4
   1.3 State of the art 6
   1.4 Research objectives 8
   1.5 Outline of this book 8

2. MEASUREMENT AND MODELING OF MISMATCH 13
   2.1 Measurement setup 14
      2.1.1 Measurement system 14
      2.1.2 Test structures 16
      2.1.3 Measurement algorithm 19
   2.2 Experimental setup 20
   2.3 Modeling of mismatch in the drain current 21
      2.3.1 Modeling approach 22
      2.3.2 Impact of threshold voltage mismatch 23
      2.3.3 Impact of current factor mismatch 28
      2.3.4 The complete model 30
      2.3.5 Parameter extraction 31
      2.3.6 Model accuracy 35
   2.4 Width and length dependence 35
      2.4.1 Width and length dependence of $\sigma^2_{\Delta P}$ 35
      2.4.2 Width and length dependence of correlation factors 37

2.4.3 Matching properties of a 0.18 μm CMOS process 38
2.5 Example: Yield of a current-steering D/A converter 41
2.5.1 Accuracy of unit current cell based on a yield requirement 42
2.5.2 Width and length of the unit current cell 42
2.6 Conclusions 45
3. PARAMETER EXTRACTION 47
3.1 Extraction methods 48
3.2 Experimental setup 51
3.3 Comparison of extraction methods 52
3.3.1 Model accuracy 53
3.3.2 Measurement accuracy and speed 55
3.3.3 Physical meaningfulness of parameters 61
3.3.4 Summary 68
3.4 Future issues 68
3.5 Conclusions 70
4. PHYSICAL ORIGINS OF MOSFET MISMATCH 73
4.1 Basic operation of the MOS transistor 74
4.1.1 Regions of operation and current expressions 74
4.1.2 Short- and narrow-channel effects 80
4.1.3 Gate depletion 85
4.1.4 Quantummechanical effects 85
4.1.5 Low field mobility 86
4.2 Mismatch in the drain current 88
4.2.1 Solution of the current equation in weak inversion 89
4.2.2 Solution of the current equation in strong inversion 94
4.2.3 Short- and narrow-channel effects 98
4.2.4 Comparison of mismatch in weak and strong inversion 101
4.2.5 Asymmetry of MOSFET mismatch 105
4.3 Physical origins of fluctuations 107
4.3.1 Doping fluctuations 108
4.3.2 Impact of fluctuations in channel doping on threshold voltage 109
4.3.3 Gate depletion 114
4.3.4 Quantummechanical effects 115

4.3.5 Mobility fluctuations 118
4.3.6 Combination of all effects and comparison with experiments 121
4.3.7 Discussion 125
4.4 Conclusions 126

5. TECHNOLOGICAL ASPECTS 129
5.1 Technology descriptions 130
5.2 Impact of the gate 133
5.2.1 Amorphous or poly-crystalline silicon as gate material? 133
5.2.2 Impact of the gate doping 136
5.3 Impact of the halo implantation 138
5.3.1 Long- and wide-channel transistors 140
5.3.2 Short- and narrow-channel effects 143
5.4 Comparison of different CMOS technologies 145
5.5 Alternative device concepts 148
5.6 Conclusions 150

6. IMPACT OF LINE-EDGE ROUGHNESS 153
6.1 Characterization of line-edge roughness 154
6.2 Modeling the impact of line-width roughness 157
6.2.1 Impact of line-width roughness on the threshold voltage 158
6.2.2 Impact of line-width roughness on the off-state current 160
6.2.3 Impact of line-width roughness on yield 162
6.3 Experimental investigation of the impact of LWR 163
6.3.1 Experimental setup 163
6.3.2 Sinusoidally-shaped gate edges 164
6.3.3 Extra rough gates 168
6.3.4 Yield 171
6.4 Prediction of the impact of LWR and guidelines 172
6.5 Conclusions 175

7. CONCLUSIONS, FUTURE WORK AND OUTLOOK 177
7.1 Conclusions 177
7.2 Future work 179
7.3 Outlook 181

Appendices 183
A List of symbols 183
B List of acronyms 189
C Publications by the author 191

References 193
About the Authors 205

Preface

This book examines the matching properties of deep sub-micron MOS transistors. Microscopic fluctuations cause stochastic parameter fluctuations that affect the accuracy of the MOSFET. For analog circuits this determines the trade-off between speed, power, accuracy and yield. Furthermore, due to the down-scaling of device dimensions, transistor mismatch has an increasing impact on digital circuits. Good insight in the magnitude of the fluctuations and their physical origins is therefore required.

This work studies the matching properties of MOSFETs at several levels of abstraction. Firstly, a simple and physics-based model is presented that accurately describes the mismatch in de drain current for the full bias range above the threshold voltage. This facilitates accurate circuit design for deep sub-micron technologies. Secondly, the most commonly used methods to extract the matching properties of a technology are bench-marked with respect to model accuracy, measurement accuracy and speed, and physical contents of the parameters. This creates insight in which method to use in which situation and in how to treat data presented in literature. As third topic the physical origins of microscopic fluctuations and how they affect MOSFET operation are investigated. This leads to a refinement of the generally applied $\sigma_{\Delta P} \propto 1/\sqrt{area}$ law in both weak and strong inversion. In addition, the analysis of simple transistor models highlights the physical mechanisms that dominate the fluctuations in the drain current and transconductance. The fourth topic considers the impact of process parameters on the matching properties. In accordance with literature, it is found that the granular structure of the poly-silicon gate material can play an important role. Furthermore, it is identified that the gate does not act as an ideal mask for the halo implantation, which worsens the matching properties of a technology. Also, scaling issues are briefly addressed. Finally, the impact of gate

line-edge roughness is investigated, which is considered to be one of the roadblocks to the further down-scaling of the MOS transistor. The impact of line-edge roughness on parameter fluctuations, off-state current and yield has been modeled. The effect has also been experimentally studied by intentionally increasing the roughness and by studying transistors with sinusoidally shaped gate edges. A prediction is made about the technology node at which line-edge roughness will become an issue. Summarizing, regarding the matching properties of deep sub-micron MOS transistors, this book tries to present insight in the modeling aspects, characterization aspects, the physical origins, and technological aspects, while also extensively treating one of the main future issues. This work could therefore be useful for device physicists, characterization engineers, technology designers, circuit designers, or anybody else interested in the stochastic properties of the MOSFET.

Acknowledgments

I would like to acknowledge the following persons for their contributions to this work.

First of all I'd like to thank coauthors prof. Herman Maes and prof. Willy Sansen. In the past few years, their thorough review of my work helped to significantly increase the quality. Dr. ir. Stefaan Decoutere, who supervised this work, is greatly acknowledged for his support, guidance, and many interesting discussions.

Special thanks must go to Hans Tuinhout from Philips Research in Eindhoven. A lot of the ideas presented in this book would not have been realized without him, especially those related to the chapter on parameter extraction. Besides Hans, I'd also like to acknowledge Régis Difrenza from ST Microelectronics in Crolles, and Johan Knol and Antoine Moonen from Philips Semiconductors in Nijmegen for their contributions to this part of the work and for fruitful discussions.

I'd also like to express my gratitude to Maarten Rosmeulen. I'm still using the environment for matching analysis that he created, and a lot of the ideas regarding the description of the mismatch in the drain current originated from him.

The work on line-edge roughness would not have been possible without the help of Peter Leunissen. I greatly appreciate our discussions on the fundamental aspects of the topic, setting up the experiments and the time he spent on creating the required test structures.

I'd also like to thank all my (former) colleagues in IMEC for making it such an inspiring environment to have worked in.

I'd finally like to acknowledge my family and friends for their contributions to life after working hours.

Jeroen Croon
November 2004

Chapter 1

INTRODUCTION

No two transistors are the same. When closely examined, differences can be observed at several levels, that, in one way or the other, are related to distance. For instance, when two 'identical' circuits are not fabricated in the same facility, they are produced by different people using different machines. This results in slightly nonidentical circuits and different circuit yields for the two different plants. In order to minimize differences, strategies like the 'copy EXACTLY! technology transfer method' of INTEL can be employed [1]. However, even within one production facility, differences between 'identical' circuits are observed. Different lots are not always processed using the same machines, while a machine itself shows a slight drift in time, which causes differences between wafers. On a single wafer, differences between dies are observed, which are called inter-die variations. These could for example be due to the fact that during processing the temperature is slightly different at the edge of a wafer than at its center.
The above effects are summarized in figure 1.1. The variation between circuits increases as their distance at process time increases. At the bottom of the upturned pyramid the intra-die fluctuations are present. Intra-die fluctuations are the differences between supposedly identical structures within one die. These differences can have a systematic nature when they are caused by asymmetries in layout. For instance, it was shown in [2] that the proximity of metal wiring lines can affect transistor operation. This e.g. reduces the mirror factor of a current mirror when one of the two transistors is more closely located to the metal line, which needs to be taken into account when the circuit is designed.
Besides systematic mismatch, also a stochastic component is present

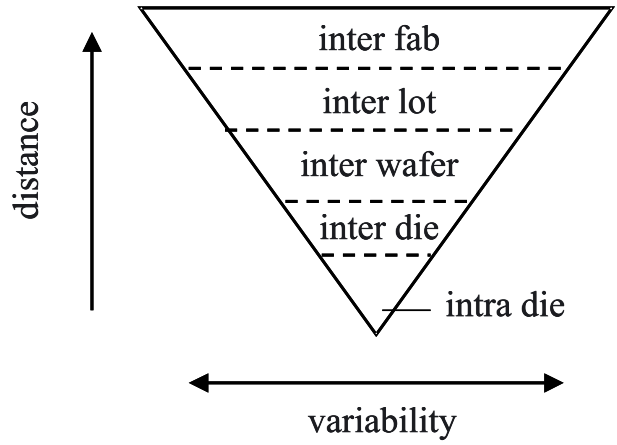


Figure 1.1. Variability at several levels

that is caused by the fact that at the microscopic level[1] transistors are not the same. One of the most well known examples of stochastic fluctuations in MOSFETs is the random nature of the amount of dopant atoms and their positions [3]. Stochastic fluctuations are independent of the distance between the devices under study, and by this they determine the maximal obtainable accuracy within a certain technology. In this work, we study the stochastic fluctuations of the MOSFET, which is the most important component of modern-day integrated circuits.

1.1 Matching analysis

The overall variability of a component is the sum of the variabilities at all levels. When studying the stochastic component, we want to filter out all other possible causes of variation. This is achieved by matching analysis, which characterizes the difference between two devices. Consider figure 1.2, which shows two types of variation: 1) Microscopic fluctuations typically have a length scale that is shorter than the device dimensions, and can be considered as spatial noise. 2) The other types of variations have length scales that are longer. Now look at the differences between the three devices that are depicted in figure 1.2. The difference between the first and third device is for the largest part due to a disturbance close to device 3, of which the impact lessens as distance increases. In other words, because the surroundings of device 1 and device 3 are nonidentical, their behavior is also nonidentical. This is often caused by

[1] Or at the nanoscale level for modern-day devices.

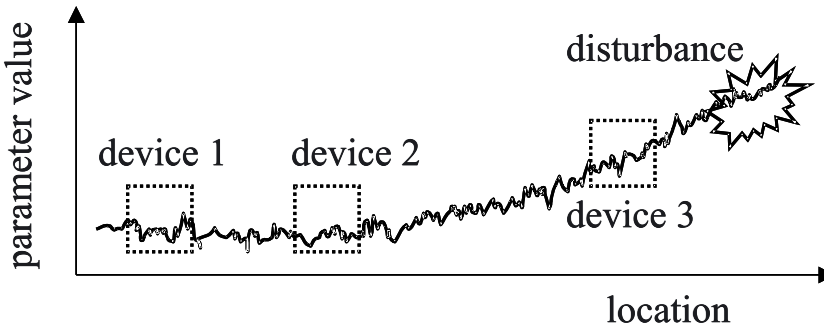


Figure 1.2. A certain device parameter as a function of the location on the chip. Devices are located at three positions.

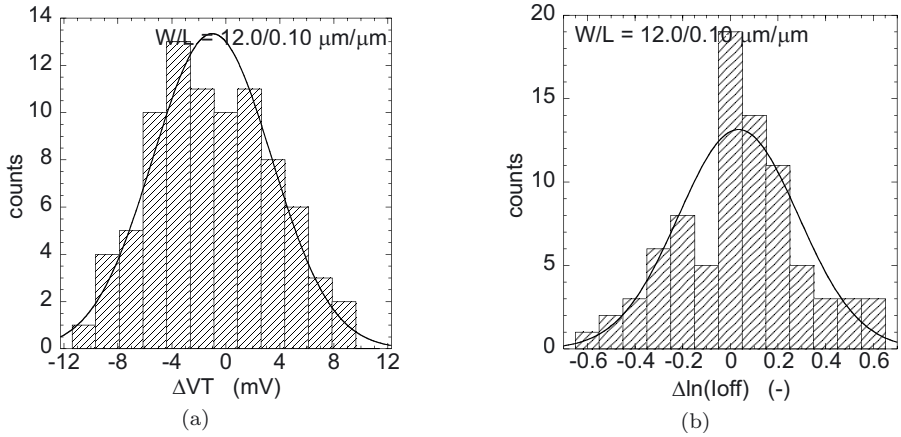


Figure 1.3. Distribution of the mismatch in the threshold voltage (a) and of the mismatch in the off-state current (b).

asymmetry in the layout, which means that the difference is systematic and the same for all processed chips. The difference between device 1 and 2 is only marginally affected by the disturbance close to device 3. Therefore, it is mainly caused by the stochastic variation. This means that the difference, or mismatch, between device 1 and device 2 is not the same as and uncorrelated to the difference observed on another chip. All this results in distributions for the mismatch as displayed in figure 1.3. Examples are shown for the mismatch in threshold voltage (ΔV_T) and the mismatch in the logarithm of the off-state current ($\Delta ln(I_{off})$). When a quantity is determined by a summation of numerous independent variables, its distribution tends to be normal, as is observed for

the mismatch in threshold voltage. The average value is determined by the systematic component of the mismatch (denoted by $\mu_{\Delta V_T}$ or $\overline{\Delta V_T}$), which is close to zero for a symmetric layout. The width of the distribution is caused by the stochastic component and it is represented by the standard deviation ($\sigma_{\Delta V_T}$) or by the variance, which is the square of the standard deviation.
Another distribution that will be encountered is the lognormal distribution, which arises when the exponent of a normally distributed parameter is taken. Lognormal distributions appear when numerous independent variables are multiplied. For the examples displayed in figure 1.3, it is observed that the mismatch in the off-state current can be approximated by such a distribution. In general, it will be found that the off-state current has a distribution in between normal and lognormal.
The difference between two devices is in most cases not represented by just one parameter. However, when more parameters are needed, these do not have to be independent from one another and correlations can exist. For instance, the off-state current is a function of the threshold voltage and a correlation between the fluctuation in these parameters can be expected.
Summarizing, when studying the matching performance of a technology, one examines the means of, standard deviations of, and correlations between the mismatch of relevant device parameters. The mismatch between two transistors increases when the distance between them is increased.

1.2 Importance for circuit design

In order to understand the impact of stochastic fluctuations, three circuit examples from literature are presented. These deal with the speed-accuracy-power trade-off in analog circuits, analog-to-digital converters, and with the SRAM circuit.
In [4] the impact of threshold-voltage mismatch on the speed-accuracy-power trade-off of analog CMOS circuits is investigated. The current mirror is examined as basic current-processing block. As basic voltage-processing block a one-transistor implementation of a voltage amplifier is taken. The size dependence of the mismatch is proportional to the inverse of the square-root of the area [5], i.e. $\sigma_{\Delta V_T} = A_{\Delta V_T}/\sqrt{area}$, where the proportionality constant $A_{\Delta V_T}$ characterizes the matching performance of a technology. Using this law, it is seen that the accuracy of a MOSFET can be increased by increasing its width or length. However, an increase in the width of a MOSFET results in a larger current and thus power dissipation. Increasing the length reduces the current, but it also reduces the speed. A similar reasoning can be applied for the

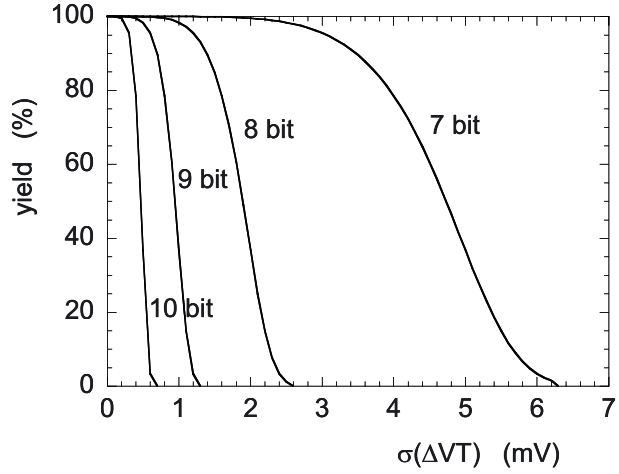


Figure 1.4. Yield of several analog-to-digital converters with different accuracies as a function of the standard deviation of the mismatch in threshold voltage. Results are taken from [6].

impact of noise. However, for the basic building blocks it is found that the impact of the matching performance of a technology on the speed-accuracy-power trade-off is one to two orders of magnitude larger than that of noise.

As second illustration, we take a look at the work presented in [6], in which the impact of stochastic variations on the yield of an analog-to-digital converter is investigated. The results of this work are copied into figure 1.4. It is indeed observed that a good matching performance is required to be able to make high accuracy analog-to-digital converters with acceptable yield.

As third example consider the SRAM circuit, which is embedded in many digital designs. Figure 1.5 shows a six transistor implementation of an SRAM cell and its transfer characteristic during read access. In [7, 8] the impact of stochastic variations in the threshold voltage on the SRAM is analyzed. This variation translates into a variation on the static noise margin (SNM), as defined in figure 1.5b. When the variation is too large, the SNM of some cells disappear, as is shown in figure 1.5 with the dashed line. In this case it is not possible to change the state of the cell and therefore it fails. It was found in [8] that in order to obtain a 90 % yield on a 1 Mbit SRAM it is required that $A_{\Delta V_T} < 6$ mVμm for a 180 nm technology and $A_{\Delta V_T} < 2.5$ mVμm for a 100 nm technology. This last number is not easy to achieve and it explains the increasing interest in research regarding stochastic parameter fluctuations.

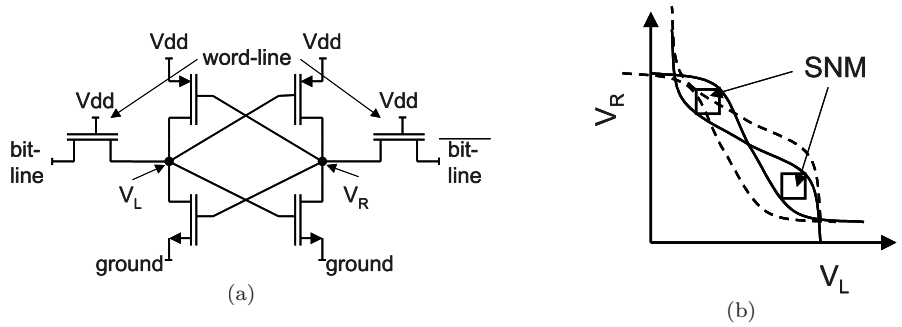


Figure 1.5. Schematic of an SRAM cell (a) and its transfer characteristic during read access (b). b) The full lines give the transfer characteristic in case of average transistor behavior. The dashed lines represent an extreme case, for which the static noise margin is reduced to zero due to the stochastic variation in the threshold voltage. The static noise margin (SNM) is equal to the length of the side of the minimum square in the 'eyes' of the transfer characteristic. These figures are based on [7].

Summarizing, it can be stated that stochastic fluctuations limit the maximal obtainable accuracy, speed, size, yield and/or minimal obtainable power dissipation in CMOS circuits.

1.3 State of the art

Looking at the references made throughout this book, it is observed that only about one third is from before 1998, which is when this work was started. This indicates the increasing interest in research regarding the matching properties of MOSFETs. Here a brief introduction is presented to the most significant papers in this field. More complete references to literature will be made at the relevant places in this book. Three kinds of topics are distinguished. The first investigates the physical origins of MOSFET mismatch, the second models the mismatch in the drain current in terms of the mismatch in other transistor parameters, and the third investigates technology related issues. Note that one publication can treat more than one of these topics.
One of the first investigated effects of microscopic fluctuations on MOSFET operation was published in 1973 by Van Overstraeten, Declerck and Broux [9]. It shows that these fluctuations need to be taken into account for accurate modeling of the weak inversion current. The first paper that examines the impact of microscopic fluctuations on the stochastic properties of macroscopic MOSFET behavior was published in 1975 by Keyes [3]. It examines the impact of the discrete character of doping on the fluctuations in the threshold voltage. This is thought to determine the

lower obtainable boundary of threshold voltage fluctuations and it is still one of the most studied effects. A popular analytical derivation based on a charge-sheet approach was presented in 1997 by Takeuchi [10] and in 1998 by Stolk [11]. The topic has also been extensively studied by device simulations (see for instance the papers of Asenov [12]). Experimental work regarding doping fluctuations was presented in the mid-nineties by Mizuno [13] and in 2000 by Tuinhout [14]. We note that, until now, calculations regarding the impact of doping fluctuations are only able to explain half of the experimentally observed fluctuations in the threshold voltage. This indicates the presence of other fluctuation mechanisms.
Another extensively studied field is how the mismatch in MOSFET parameters translates into a mismatch in the drain current. In general, this is achieved by first order sensitivity analysis on a relatively simple model for the drain current. In most cases mismatch in the threshold voltage and mismatch in the current factor are taken into account (see for example the work of Vittoz [15] (1985), Lakshmikumar [16] (1986), Pelgrom [5] (1989), Bastos [17] (1995) and Serrano-Gotarredona [18] (2000)). Drennan [19] (1999) follows a slightly different approach by starting from a more complex compact model and by assuming prior knowledge of width and length dependencies to estimate model parameters.
Maybe the most referred to paper in matching literature is the one written in 1989 by Pelgrom [5]. This work examines the width and length dependence of the standard deviation of the mismatch at the fundamental level. This standard deviation is found to be inversely proportional to the square root of the device area. This is one of the best known laws in the field of matching.
The impact of technology-related parameters is less well understood. However, some effects were studied, like for instance the influence of metal coverage [20] (1996) and the impact of the granular structure of the gate material [21] (1997) by Tuinhout. The impact of the vertical doping profile was studied by Takeuchi [10] (1997), while e.g. Difrenza looked at the impact of halos [22] (2000). In 2001 Stolk [8] briefly outlined the required steps to optimize a technology with respect to its matching performance. However, note that technologies keep changing and that this work can never be considered complete.
Summarizing, we conclude that research of the stochastic properties of technologies is gaining in interest. Knowledge has been built up regarding the impact of doping fluctuations on the threshold voltage and of how the mismatch in the drain current depends on transistor parameters. Technology-related issues have been investigated, but are not completely understood. Furthermore, with the down-scaling to deep submicron and

sub 100 nm gate lengths, new technological and physical issues arise. In general it can be stated that full quantitative understanding of the matching properties of MOSFETs is still missing.
Finally, references should be made to the Ph.D. theses of Bastos [23] (1998), Difrenza [24] (2002) and Tuinhout, that deal with the topic of matching. Bastos mainly concentrated on the description of the mismatch in the drain current and on the impact of mismatch on a digital-to-analog converter. Difrenza focussed on the physical modeling and also discussed the impact of the gate material and the halo implantation. Based on numerous practical examples, Tuinhout extensively studied the measurement of mismatch and layout issues.

1.4 Research objectives

The main goal of this work is to understand, model and characterize the matching properties of deep submicron MOSFETs. This is further specified as:

- Develop a physics-based model that accurately describes the mismatch in the drain current over as large a bias range as possible.
- Benchmark different methods for mismatch characterization.
- Understand and provide models for the physical causes of MOSFET mismatch.
- Investigate the impact of process steps and technological parameters on the matching performance of deep-submicron technologies.
- Investigate the impact of line-edge roughness as one of the future causes for stochastic parameter fluctuations.

These objectives encompass all three matching research topics defined in the previous subsection. The work presented in this book is done on 180 nm and 130 nm CMOS technologies developed in IMEC. Experimentally investigated gate lengths range down to sub 100 nm.

1.5 Outline of this book

This book consists of five technical chapters after which it is concluded and suggestions for future work are presented. The chapters are related to the above mentioned research objectives and are presented in the same order. This also approximates the chronological order in which the work took place. Exceptions are chapter 4, for which the work was done last, and chapter 5, which shows results that were obtained during the full duration of this work.

We started our work in 1998 in IMEC by trying to describe the mismatch in the drain current as a function of other model parameters (chapter 2). We reasoned that, by taking a physical model as base, this would automatically lead to physical insight in the matching properties of the MOSFET. This turned out to be only partly true. By the time the work for chapter 2 got finalized, Philips Research and IMEC had started working together. Comparison of our extraction methodology with the one of Philips uncovered large and unexpected differences. This resulted in a small collaboration between Philips Research (Eindhoven, the Netherland), Philips Semiconductors (Nijmegen, the Netherlands), ST Microelectronics (Crolles, France) and IMEC. The same material was measured at each of these locations and the most common extraction methods were bench-marked. The results of this work are presented in chapter 3. By now it became apparent that a deeper knowledge regarding the physical origins of MOSFET mismatch was required, and a lot of the ideas that ended up in chapter 4 were developed in this period. At the same time, in the lithography group of IMEC the question arose how to deal with line-edge roughness. Another small collaboration was started, and priority was given to this work. However, some of the ideas regarding the physical origins of mismatch could already be applied for the specific case of line-edge roughness. We have therefore chosen to present the work regarding line-edge roughness in chapter 6 at the end of this book as an illustration of the more general theories presented earlier. By now our ideas regarding the physical origins of MOSFET mismatch had received time to mature. They are presented in chapter 4. In order to understand technological issues, relevant process splits were analyzed during the full duration of this work. Also, a dedicated experiment was set up, which mainly focussed on the impact of the halo implantation. The results of this work are presented in chapter 5.
A more detailed overview of the contents of the chapters will now be given.

Chapter 2: Measurement and modeling of mismatch in the drain current. The main topic of this chapter is the modeling of the mismatch in the drain current as a function of mismatch in the threshold voltage and current factor. An accurate model is required in order to fully understand the impact of variability on the MOSFET and to evaluate the impact of mismatch on circuits. We distinguish ourself from other work by our modeling approach: The impact of the mismatch in threshold voltage and current factor are treated separately. Assumptions that are required to model the impact of mismatch in the current factor are not required to model the impact of mismatch in the threshold voltage. This approach results in a continuous model that is valid in moderate

and strong inversion.
Most of the theories presented in this book are compared to experimental data. Therefore, chapter 2 starts by describing our measurement setup, test structures and measurement approach.

Chapter 3: Parameter extraction. Numerous methods exist that extract the variation in the threshold voltage and current factor. Quite often publications do not mention which method is used, but we will show that significant differences can occur. The most commonly applied methods are bench-marked with respect to model accuracy, physical meaningfulness, and measurement accuracy and speed. The following methods are examined: the maximum slope method, the three points method, the four points method, applying a current criterion and current-mismatch fitting methods.

Chapter 4: Physical origins of MOSFET mismatch. This chapter looks at the origins of fluctuations at the microscopic level and at how they affect MOSFET behavior. In order to achieve this, it is necessary to delve deeper into MOSFET theory than before and, as an introduction to this chapter, the basic equations of MOSFET operation are derived. The chapter continues by again deriving these equations, but now in the presence of microscopic fluctuations. In agreement with other published work, we find that the $1/\sqrt{area}$ law does not hold in weak inversion. Furthermore, we find in this regime of operation that edge effects, like halos or shallow trench isolation, can cause serious increases in the mismatch for long and wide transistors, which are not observed in strong inversion. In parallel and in agreement with a recent publication [25] we also find a slight departure of the $1/\sqrt{area}$ law in strong inversion for high enough values of the drain bias. Short- and narrow-channel effects are described using theories published in literature.
The chapter ends by using the theory of MOSFET operation to calculate the impact of doping fluctuations in the channel region and gate, the impact of fluctuations in the oxide charge and the impact of fluctuations in surface roughness. As in literature, the charge sheet approach is followed. The calculations include quantum mechanical effects, gate depletion and fluctuations in the mobility. We predict that Coulomb scattering gives a significant contribution to stochastic parameter fluctuations. We combine all models and fit the total model to the experimentally obtained curve of the mismatch in the drain current as a function of the gate bias. The physical content of the model is tested by predicting the mismatch in the transconductance, the mismatch at different bulk bias conditions, and the correlation of the mismatches at several bias conditions.

Chapter 5: Technological aspects. In this chapter examples are presented that demonstrate how certain process parameters can affect the

matching properties of a technology. As in literature we find that the grain structure of the gate material can have a large impact. Furthermore, the impact of the halo implantation is examined. We find that halos can seriously degrade the matching performance of a technology when they are unintentionally implanted through the gate. Also in this chapter, the scaling behavior of the matching performance is addressed.

Chapter 6: Impact of line-edge roughness on parameter fluctuations, off-state current and yield. For near-future gate-lengths, line-edge roughness is expected to cause significant parameter fluctuations, increase the off-state current and decrease yield. Therefore, it has recently become a topic of interest. The chapter starts with the description of line-edge roughness itself. Based on this information, we calculate the impact of line-edge roughness. We test our models by intentionally increasing the roughness. We then use these models to predict the moment at which line-edge roughness will become an issue. These predictions are used to present guidelines for as well device engineering as gate-patterning process development.

This book ends in *chapter 7* with the major conclusions and suggestions for future work.

Chapter 2

MEASUREMENT AND MODELING OF MISMATCH IN THE DRAIN CURRENT

In order to be able to calculate or simulate the effects of MOS transistor mismatch, it is important to have a model that accurately describes the mismatch in the drain current. In the development of such a model, several aspects have to be taken into account. We would like the model to be *valid over a large bias range*. This would allow us to use the model for a large set of applications. We would also like the model to be *physics based*. A physics based model has as advantage over empirical models, that its model parameters can more easily be linked to the technology of which it describes the matching properties. Thirdly, the model needs to be *continuous* between different regions of operation of the MOS transistors. Continuity makes the model easier to implement in a circuit simulator. Furthermore, a method needs to be developed to *extract model parameters*.

Existing physics based mismatch models[1] can be separated in two groups, those that take a complex description of the drain current as base [19, 28–30], or those based on a simple description [5, 15–18, 31–41]. The mismatch models based on complex drain-current models can provide very accurate results. They include a lot of well understood physical effects and therefore contain many parameters. However, the mechanisms that cause mismatch are usually only partly understood and one cannot implicitly assume that a mismatch model automatically inherits the physics contained in the drain current model on which it is based. E.g., although a parameter is independent of a certain bias voltage, this does not have to hold for the mismatch in that parameter.

[1] As opposed to statistical modeling [26, 27].

Because of the large number of parameters involved, complex models are very time consuming to use.
In this chapter, we therefore choose to develop a mismatch model that yields sufficiently accurate results, but that is kept as simple as possible (section 2.3). Our newly developed modeling approach has the advantage over previously published work that it stretches the range of validity towards lower values of the gate bias into the upper part of the moderate inversion region. Also the bias dependence of the mismatch parameters will be investigated. A new scheme for parameter extraction will be introduced. The width and length dependence of the extracted parameters is described by the model published in [5], which is presented in section 2.4. As an illustration, in section 2.5 the model is used to determine the width and length of the current-source transistor in the unit current cell of a current-steering digital-to-analog converter.
The model, derived in this chapter, will be tested on a 0.18 μm CMOS technology, from which measurement data is required. We will therefore start by describing how to measure MOS transistor mismatch (section 2.1) and by providing the experimental background (section 2.2). Section 2.6 concludes the chapter.

2.1 Measurement setup

In order to determine the mismatch between two transistors, we want to measure their drain currents as simultaneously as possible. This requires an appropriate measurement system, test structures and measurement algorithm. These issues will be discussed in the following three subsections, respectively.

2.1.1 Measurement system

The system used for the measurements is schematically presented in figure 2.1. This system is part of the semi-automatic HP4063 Semiconductor Parameter Analysis System, which is described fully in [42]. It consists of a wafer-prober, a chuck, a switching matrix, a parameter analyzer and an UNIX workstation. Also needed is a probe-card. These component will now briefly be described.

Wafer-prober. Accurate extraction of a standard deviation requires the measurement of a large number of transistor pairs. To measure manually would therefore be a very time consuming and tedious process. The wafer-prober automatically moves the chuck around so that all required device modules on the wafer are contacted.

Chuck. The measurement-wafer is located on a thermochuck. The operation of the thermochuck induces noise in the transistors under test,

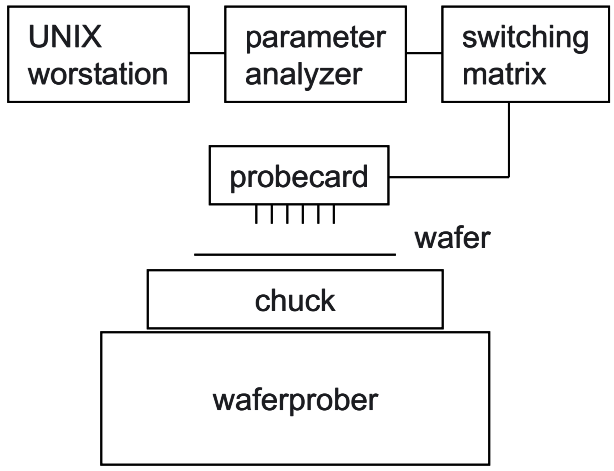


Figure 2.1. Measurement system

which can seriously degrade measurement accuracy. The two transistors of a pair are measured directly after each other, while the temperature difference between two consequential measurements is not significant. Therefore, the thermochuck is switched off.

Switching matrix and probe-card. One module, which consists of 2×12 bonding pads, contains several transistor pairs. All twenty-four bonding pads are contacted at once by the twenty-four pins of the probe-card. The type of probe-card used depends on the material of the bonding pads and is chosen in such a way as to minimize the contact resistance. For aluminum bonding pads a probe-card with tungsten needle-tips is used, while for copper bonding pads the needle tips are made of a beryllium-copper alloy. Although the probe-card has twenty-four pins, the parameter analyzer only has four SMUs. The switching matrix takes care of connecting the correct pin to the correct SMU.

Parameter analyzer. Through its four SMUs, the HP4142B parameter analyzer supplies the bias voltages and measures the currents of the transistors under test. A force and sense technique is applied for the biasing. The sensing is done in between the probe-card and the switching matrix. For most of the measurements only the lowest voltage range ($-2 < V < 2$) of the system is needed, which has a more than sufficient resolution of 100 μV. The specified worst case accuracy is < 2.1 mV. The resolution at which currents are measured is 0.02 % at the bottom of a specific measurement range and 0.002 % at the top of the range. Changes in measurement range take place at current levels of approximately 10^n A, where n is an integer. The specified worst case accuracy is

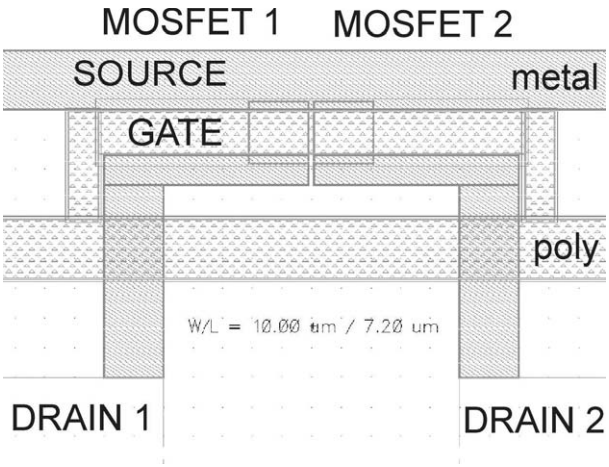


Figure 2.2. Layout of standard transistor pair

~ 0.5 % of the measured current. When doing matching measurements the non-specified short-term repeatability of the measurement system is of far more importance than the worst case accuracy. This will be extensively analyzed in section 3.3.2, where we will find that this measurement repeatability is much better than the specified worst case accuracy.

UNIX workstation. The UNIX workstation is used to communicate with the measurement equipment and to collect the measurement data.

2.1.2 Test structures

This section introduces the test structures that are needed to characterize MOS transistor mismatch. A nice overview of test structures for matching studies was published in [43]. Figures 2.2 and 2.3a display the standard matched transistor pair. The two transistors have common gate, common source and common bulk. Their drains are connected separately. With the standard test structure we only want to analyze random local fluctuations. Therefore the test structure is designed to be as symmetrical as possible, the transistors are located close to each other and their currents flow in the same direction.
Mismatch can also be due to systematic differences in layout or by longer range gradients. To analyze this kind of mismatch, different test structures are required. We will discuss the most common ones, which are also presented in figure 2.3. Note that, according to need, numerous kinds of variations to these test structures are possible.

Rotated transistors. Differences in e.g. crystal orientation and stress can cause systematic mismatch between transistors with different orien-

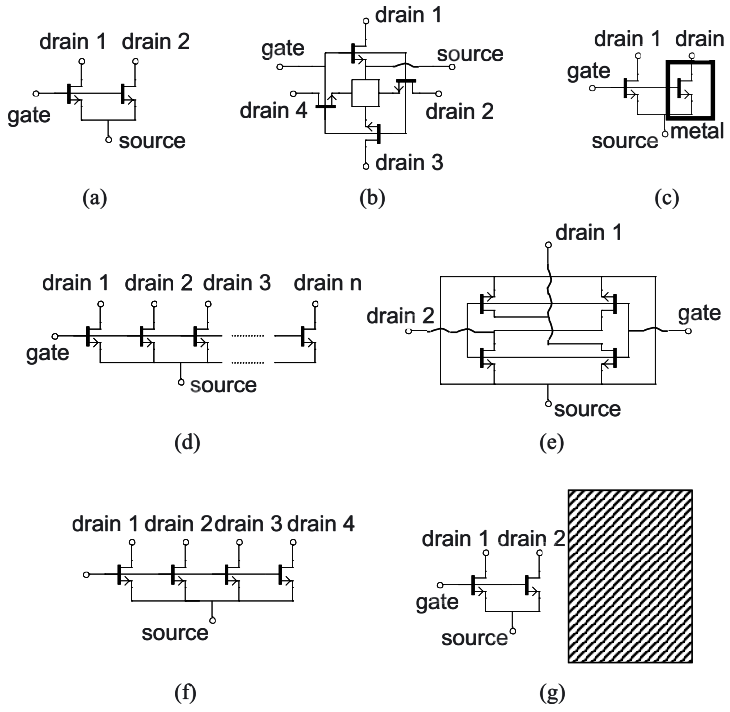


Figure 2.3. Schematic representation of the standard matched transistor pair (a) and test structures for evaluating the mismatch between transistors with different orientations (b), the impact of metal coverage (c), gradient mismatch (d), a quadrature layout (e), the influence of dummy transistors (f) and the impact of the proximity of a large structure like a resistor or capacitor (g)

tation [44]. A test-structure in which transistors are rotated with respect to each other allows for investigation of these effects.

Metal coverage. For easy routing of metal lines it would be favorable if they could be lain-out over transistors. This can cause systematic deviations due to e.g. insufficient passivation of dangling bonds at the silicon silicon-dioxide interface [20, 45, 46]. To investigate the influence of the proximity of a metal line, a transistor pair is designed in which one of the transistors is covered with metal.

Impact of gradients. Layer thicknesses and doping profiles can vary slightly over a chip or wafer. These gradients cause a systematic mismatch, which becomes more prominent when two transistors are located further apart. As test-structure an array of transistors is used, which are spaced at a certain distance.

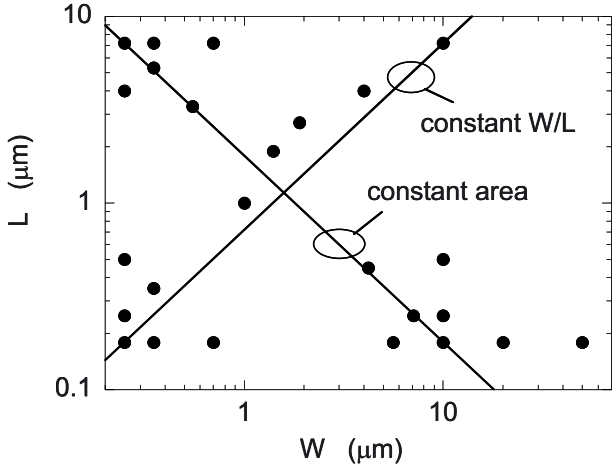


Figure 2.4. Transistor pair dimensions for the evaluation of random mismatch for a 0.18 μm CMOS process

Quadrature layout. A way to circumvent the effects of gradients or other systematic mismatch causing effects is to use a quadrature layout. Each transistor in the pair is split up in two transistors. The four resulting transistors are cross-coupled (see figure 2.3e). This structure is quite complex to lay out, which might actually result in extra mismatch contributions.

Dummy transistors. When several transistors are supposed to match, the transistors at the side of an an array have different surroundings than transistors in the center. This can cause systematic mismatch. Adding dummy transistors at the edge of the array reduces this mismatch. As test structure, several closely spaced transistors are placed next to each other. All terminals are common, except the drain connections.

Impact of a capacitor or resistor. The proximity of a large structure, like a capacitor [47] or resistor, can affect transistor behavior and cause mismatch. As test-structure, matched pairs are placed at several distances from the large structure under investigation.

We will now return to the standard matched transistor pair, used for extracting random mismatch. In section 2.4 it will be found that random mismatch is inversely proportional to the square root of the area. Deviations are expected for short or narrow transistors. To examine this width and length dependence a proper set of dimensions needs to be defined. As an example, figure 2.4 shows the chosen dimensions for a test-chip of a 0.18 μm technology. In this figure, transistors on the diagonal going from lower-left to upper-right have constant W/L-ratio, but different

areas. These pairs are used to evaluate the area dependence. Transistors on the other diagonal have constant area, but different W/L-ratios. They are used to evaluate the impact of short- and narrow-channel effects.
Having chosen the device dimensions, another issue that needs to be dealt with is the total required amount of transistor pairs with the same dimension (N_{dev}). Usually, the measured mismatch in the drain current can be assumed normally distributed. Then, from basic statistical theory, it follows that the standard deviation (σ_σ) of the extracted standard deviation (σ) is equal to:

$$\sigma_\sigma = \frac{\sigma}{\sqrt{2N_{dev}}}. \tag{2.1}$$

One 8" wafer contains approximately 40 to 200 test chips. One experiment of ∼20 wafers, usually has 2 wafers per experimental split. With one transistor pair per dimension per test-chip, this gives rise to σ_σs ranging from 3.5 % to 8 %, which is sufficiently accurate for most experiments.

2.1.3 Measurement algorithm

This subsection describes the routine, which is used to measure the drain currents (I_D) of the two transistors of the pair under test. MOS transistor mismatch is usually evaluated as a function of the gate bias (V_{GS}) at a certain drain bias (V_{DS}) and bulk bias (V_{BS}). The gate bias ranges from 0 V to the supply voltage (V_{DD}). Steps of 50 mV are sufficiently small. Because conditions of the surroundings (e.g. temperature) can vary over time, we want to measure these curves as fast 'after' each other as possible. This is done in the following way. First the voltages are supplied to the common source, common bulk, common gate and separate drains. Next the drain current of the first transistor is measured, then the drain current of the second transistor. The gate bias is increased (or decreased in the case of PMOSFETs) by 50 mV and again the two drain currents are measured directly after each other. This process is repeated until the full $I_D - V_{GS}$ curves are measured. Note that our main interest lies in the difference between parameters. The absolute measurement conditions are therefore not of great importance, as long as these conditions are stable.
The drains of the two transistors are routed to two separate SMUs. This might give rise to a measurement-system-related offset. To circumvent this problem the measurement is repeated, but transistor one is now considered as the second and vice versa. Combining the two measurements cancels out the offset. As mentioned before, another source of

error could be the variation in temperature during the measurement. However, in the next chapter a good measurement repeatability will be demonstrated, and it is thus concluded that temperature fluctuations do not play a significant role.
As mentioned in the previous subsection, accurate extraction of standard deviations requires a lot of measurements. This makes measurement time a serious constraint. The algorithm presented above needs approximately 150 ms to measure one bias condition. Repeating the measurement makes this 300 ms. As practical example we will take the measurement of a 0.18 μm technology with a supply voltage of 1.8 V. When 15 pair dimensions are examined at two values of the drain bias, one wafer contains 40 chips and both NMOS and PMOS transistors are measured, the total measurement time for one wafer would approximately be $7\frac{1}{2}$ hours.

2.2 Experimental setup

In the next section measurements are performed to test the mismatch model under development. The experimental background for these experiments will now be provided. Choices need to be made concerning: used technology, type of transistors, geometries of examined device pairs, the number of measured pairs and what to measure.

Technology. The technology chosen for this experiment is the 0.18 μm CMOS technology published in [48], which has a physical oxide thickness of 2.8 nm and a supply voltage of $V_{DD} = 1.8$ V. At the time of this research, to our knowledge, simple mismatch models had not been demonstrated on technologies with gate lengths below 0.7 μm.

Type of transistors. Both NMOS and PMOS transistors are examined. Since no significant differences were observed, most of the shown results are for NMOS transistors.

Device pair geometries. In the standard lay-out, device pairs with 25 different geometries are available on the test chip used for the experiment. The dimensions are shown in figure 2.4. To limit measurement time only the subset of 14 pair dimensions listed in table 2.1 is measured.

This subset contains approximately square transistors with different areas (left column) and transistors with constant area, but different width-over-length ratios (right column). Quite often, only results for the four emphasized geometries are shown, in order to keep the number of presented figures under control. These geometries are representative for the whole set of measured pair dimensions.

Number of measured pairs. The sample size for this experiment is 84 device pairs per pair geometry. From (2.1) it follows that this results in

Table 2.1. Measured pair dimensions. The main focus is put on the highlighted geometries.

'square'		constant area	
W (μm)	L (μm)	W (μm)	L (μm)
0.25	**0.18**	**10.0**	**0.18**
0.25	0.25	7.1	0.25
0.35	0.35	4.2	0.45
1.0	**1.0**	0.55	3.3
1.4	1.9	0.35	5.3
1.9	2.7	**0.25**	**7.2**
4.0	4.0		
10.0	7.2		

a relative accuracy of the extracted standard deviations of $\sigma_{\sigma\Delta P}/\sigma_{\Delta P} = 7.7\%$.

What to measure. To test the model, eight $I_D - V_{GS}$ curves per transistor in the pair are measured by the routine described in subsection 2.1.3. The bias conditions of the measurements are presented in the table below.

V_{DS} (V)	V_{BS} (V)	V_{DS} (V)	V_{BS} (V)
0.05	0.0	0.05	-0.9
0.3	0.0	1.8	-0.9
0.9	0.0	0.05	-1.8
1.8	0.0	1.8	-1.8

2.3 Modeling of mismatch in the drain current

In this section a model is developed that describes the relative mismatch in the drain current ($\Delta I_D/I_D$) as function of the bias voltages (V_{GS}, V_{DS} and V_{BS}). As was mentioned in the introduction of this chapter we want the model to be physics based, valid over a large bias range, continuous between different regions of operation, and as simple as possible, while sufficiently accurate. The accuracy target is:

$$\left| \frac{\sigma_{\Delta I_D/I_D}|_{model}}{\sigma_{\Delta I_D/I_D}|_{experimental}} - 1 \right| < 20\ \%. \tag{2.2}$$

This section is organized as follows. In the first subsection the applied modeling approach will be introduced. Subsections 2.3.2 and 2.3.3 calculate the impact on the drain current of a mismatch in threshold voltage and current factor, respectively. The method for parameter extraction is developed in subsection 2.3.5. Finally, in subsection 2.3.6 the model accuracy is examined.

2.3.1 Modeling approach

When modeling mismatch it can safely be assumed that the mismatch in a certain parameter ΔP is much smaller than the value of the parameter P itself. In this case the impact of the mismatch in parameters P_i on the drain current I_D can be calculated by a first order Taylor approximation:

$$\frac{\Delta I_D}{I_D} \cong \frac{1}{I_D}\frac{\partial I_D}{\partial P_1}\Delta P_1 + \frac{1}{I_D}\frac{\partial I_D}{\partial P_2}\Delta P_2 + \dots . \tag{2.3}$$

The mismatch $\Delta I_D/I_D$ in a transistor pair is just one realization of a distribution of possible $\Delta I_D/I_D$'s. This distribution can usually be assumed normal, in which case it is fully described by a mean ($\mu_{\Delta I_D/I_D}$) and a standard deviation ($\sigma_{\Delta I_D/I_D}$). From (2.3) it directly follows that:

$$\mu_{\Delta I_D/I_D} = \frac{1}{I_D}\frac{\partial I_D}{\partial P_1}\mu_{\Delta P_1} + \frac{1}{I_D}\frac{\partial I_D}{\partial P_2}\mu_{\Delta P_2} + \dots \text{ and} \tag{2.4}$$

$$\sigma^2_{\Delta I_D/I_D} = \left(\frac{1}{I_D}\frac{\partial I_D}{\partial P_1}\right)^2 \sigma^2_{\Delta P_1} + \left(\frac{1}{I_D}\frac{\partial I_D}{\partial P_2}\right)^2 \sigma^2_{\Delta P_2} + \tag{2.5}$$

$$+ \frac{2}{I_D^2}\frac{\partial I_D}{\partial P_1}\frac{\partial I_D}{\partial P_2}\rho(\Delta P_1, \Delta P_2)\sigma_{\Delta P_1}\sigma_{\Delta P_2} + \dots ,$$

where $\mu_{\Delta P}$ is the mean of ΔP, $\sigma_{\Delta P}$ its standard deviation, and $\rho(\Delta P_1, \Delta P_2)$ the correlation between the mismatches in parameters P_1 and P_2.

In accordance with previous work [5, 15–18, 28, 29, 31, 33–40], the mismatch in the drain current is assumed to result from a mismatch in threshold voltage (ΔV_T) and a mismatch in the current factor ($\Delta\beta/\beta$). Using the equations above, their impacts will be calculated in the next two subsections, respectively.

Although our model will also be based on assumptions concerning the drain current model, as apposed to other models, we will look for each parameter separately which assumptions are required. In other words, in developing our mismatch model, we did not limit ourselves to just one description of the drain current. In this way we hope to keep the

model as simple as possible. The drain current is only modeled to such an extent as is necessary to describe the mismatch, related to either the threshold voltage or the current factor.

2.3.2 Impact of threshold voltage mismatch

In calculating the impact of threshold voltage mismatch on the drain current, it is assumed that the drain current is a function of the gate-overdrive voltage ($V_{GS} - V_T$), but not of V_{GS} or V_T separately:

$$I_D = \mathrm{f}(V_{GS} - V_T, V_{DS}, V_{BS}). \tag{2.6}$$

Using (2.3), it follows that:

$$\left.\frac{\Delta I_D}{I_D}\right|_{\Delta V_T} = \frac{1}{I_D}\frac{\partial I_D}{\partial V_T}\Delta V_T \cong -\frac{1}{I_D}\frac{dI_D}{dV_{GS}}\Delta V_T = -\frac{g_m}{I_D}\Delta V_T, \tag{2.7}$$

where g_m is the transconductance. As opposed to other models we do not proceed with modeling g_m/I_D. Further working out of the term at this stage would require more assumptions and would make this part of the mismatch model unnecessarily complex. For practical applications a circuit designer can calculate g_m/I_D from any suitable drain current model. When extracting model parameters or evaluating model accuracy, g_m/I_D can be calculated directly from the measurement data. This is the approach followed in the remainder of this section.
Since (2.7) is only based on assumption (2.6) we expect this equation to be valid in the whole inversion region. The validity of this statement will now be examined. In strong inversion ($V_{GS} \gg V_T$) it approximately holds that:

$$g_m/I_D \propto 1/(V_{GS} - V_T). \tag{2.8}$$

In weak inversion ($V_{GS} \ll V_T$) the drain current can be written as[2]:

$$I_D = \frac{W}{L} I_0 e^{(V_{GS}-V_T)/n\phi_t}\left(1 - e^{-V_{DS}/\phi_t}\right), \tag{2.9}$$

where I_0 is the normalized current extrapolated to $V_{GS} = V_T$, L and W are the transistor length and width, ϕ_t is the thermal voltage kT/q and $n\phi_t$ is the subthreshold slope. From (2.8) it follows that threshold voltage mismatch becomes the dominant mismatch causing effect at low gate biases in the strong inversion region. Since in weak inversion, the drain current depends exponentially on the threshold voltage, its mismatch is

[2]All equations written down in this book are valid for NMOS transistors. The equations for PMOS transistors are easily found by introducing the appropriate minus signs.

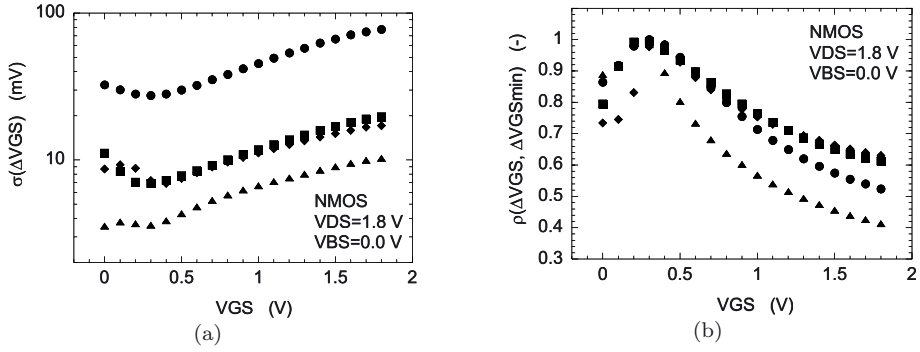


Figure 2.5. a) Experimental $\sigma_{\Delta V_{GS}} - V_{GS}$ curves. b) Correlation of ΔV_{GS} with the ΔV_{GS} value at the minimum of the $\sigma_{\Delta V_{GS}} - V_{GS}$ curve. (●) W=0.25 μm, L=0.18 μm, (■) W=10.0 μm, L=0.18 μm, (⧫) W=1.0 μm, L=1.0 μm, (▲) W=0.25 μm, L=7.2 μm

also expected to be primarily determined by threshold voltage fluctuations. In this case, it follows from (2.7) that $\sigma_{\Delta V_{GS}} = \sigma_{\Delta I_D}/g_m = \sigma_{\Delta V_T}$. The mismatch in the gate bias is evaluated at constant drain current. For several dimensions, figure 2.5a shows $\sigma_{\Delta V_{GS}}$ as a function of the average gate bias at this current. At high gate biases $\sigma_{\Delta V_{GS}}$ is increasing with V_{GS}, which suggests the dominance of current-factor mismatch. At lower gate biases the curves are expected to level off at $\sigma_{\Delta V_T}$. However, in contradiction to the observations reported in [49, 50], this behavior is not encountered. Figure 2.5 plots the correlation of ΔV_{GS} at the minimum of the $\sigma_{\Delta V_{GS}} - V_{GS}$ curve and $\Delta V_{GS}(V_{GS})$ as a function of the gate bias. At high gate bias the correlation decreases, because current-factor mismatch takes over from threshold-voltage mismatch. However, it is seen that the correlation also drops when going into the weak inversion region. The behavior observed in figure 2.5 might be due to a couple of reasons. For instance, the mismatch in threshold voltage itself can originate from different physical effects in weak and strong inversion. The difference between weak and strong inversion will be extensively studied in chapter 4, section 4.2. For the model developed in this section we conclude that (2.7) is valid for gate biases higher than the minimum of the $\sigma_{\Delta V_{GS}} - V_{GS}$ curve, which approximately lies at $V_{GS} = V_T$ (~ 0.35 V). Note that this range of validity is significantly larger than that of other simple models in literature.

Now the dependence of threshold voltage mismatch on the drain and bulk bias will be investigated. We will start with the drain bias dependence. For short transistors it has been reported that threshold

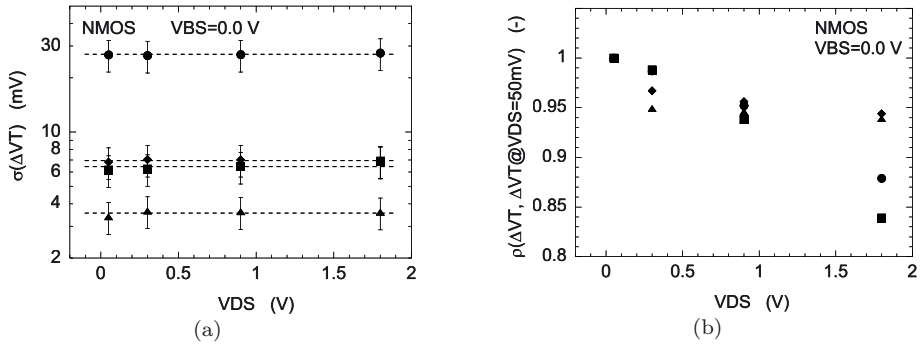


Figure 2.6. $\sigma_{\Delta V_T}$ (a) and the correlation of ΔV_T with $\Delta V_T@V_{DS} = 50$ mV (b) as a function of the drain bias. Error bars represent 99 % confidence intervals. (●) W=0.25 μm, L=0.18 μm, (■) W=10.0 μm, L=0.18 μm, (♦) W=1.0 μm, L=1.0 μm, (▲) W=0.25 μm, L=7.2 μm

voltage mismatch can increase due to drain induced barrier lowering[3] [51]. Figure 2.6a shows $\sigma_{\Delta V_T}$ as a function of the drain bias for several pair dimensions. Figure 2.6b shows the correlation of $\Delta V_T(V_{DS})$ with $\Delta V_T@V_{DS} = 50$ mV. In these figures, threshold voltage mismatch is extracted by applying a current criterion[4]. It is observed that $\sigma_{\Delta V_T}$ does not vary significantly with the drain bias. The correlation drops slightly with increased drain bias, which is more prominent for short transistors, as expected. However, since these effects are not very strong, they will be neglected.

We will continue with the modeling of the bulk bias dependence of threshold voltage mismatch. The threshold voltage can be written as:

$$V_T = V_{T0} + \gamma(\sqrt{\phi_B - V_{BS}} - \sqrt{\phi_B}), \tag{2.10}$$

$$V_{T0} = V_{FB} + \phi_B + \gamma\sqrt{\phi_B}. \tag{2.11}$$

V_{FB} is the flat-band voltage, V_{T0} is the threshold voltage at zero bulk bias, γ is the body-effect coefficient and ϕ_B is the surface potential in strong inversion. In literature, threshold voltage mismatch is usually described by a mismatch in V_{T0} and a mismatch in γ. We will not follow this approach for the following reason. The body-effect coefficient is given by:

$$\gamma = \frac{\sqrt{2q\epsilon_{si}N_A}}{C_{ox}}, \tag{2.12}$$

[3]For a physical explanation we again refer to chapter 4
[4]An overview of extraction methods will be presented in chapter 3

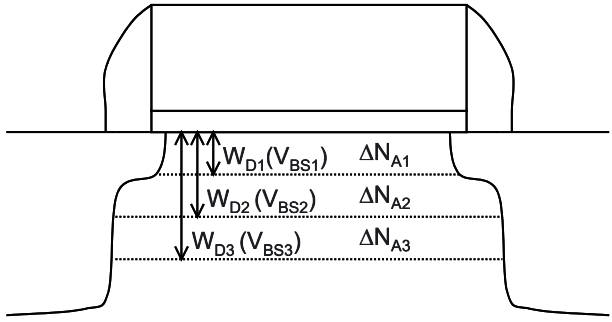


Figure 2.7. Schematic representation of the MOS transistor. The width of the depletion region W_D is drawn for three different values of the bulk bias.

where C_{ox} is the oxide capacitance per unit area, N_A is the doping concentration, q is the elementary charge and ϵ_{si} is the permittivity of silicon. Fluctuations in the threshold voltage can be attributed to fluctuations in doping concentration and fluctuations in oxide capacitance. Now consider figure 2.7, which shows a schematic drawing of the MOS transistor. The width of the depletion region is drawn for three different values of the bulk bias and $V_{BS1} > V_{BS2} > V_{BS3}$. When the bulk bias becomes more negative, the depletion width increases. The extra amount of dopants included in the depletion region (N_{A2}) fluctuates independently from the original amount of dopants (N_{A1}). In other words, the correlation between ΔN_{A1} and ΔN_{A2} is zero. The same holds for the extra included dopants when the bulk bias is decreased even further. It follows that, although for uniform doping profiles γ is independent from the bulk bias, $\Delta\gamma$ cannot be considered constant. To avoid this problem, we choose to model the bulk bias dependence of $\sigma_{\Delta V_T}$ instead of the bulk bias dependence of ΔV_T. In [11][5] the impact of doping fluctuations on threshold voltage mismatch is calculated to be:

$$\sigma^2_{\Delta V_{T,doping}} = \frac{t^2_{ox}\sqrt{8q^3\epsilon_{si}N_A(\phi_B - V_{BS})}}{3WL\epsilon^2_{ox}}, \tag{2.13}$$

In case of fluctuations in the oxide capacitance ($\sigma_{C_{ox}}$), it follows from (2.3) and (2.10) to (2.12) that:

$$\sigma_{\Delta V_{T,C_{ox}}} = \gamma\sqrt{\phi_B - V_{BS}} \cdot \frac{\sigma_{\Delta C_{ox}}}{C_{ox}}. \tag{2.14}$$

[5] Also, see chapter 4, subsection 4.3.2.

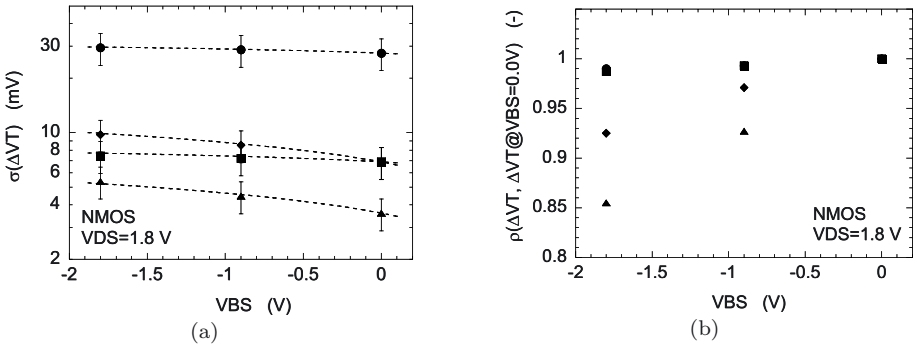


Figure 2.8. $\sigma_{\Delta V_T}$ (a) and the correlation of ΔV_T with $\Delta V_T@V_{BS} = 0$ V (b) as a function of the bulk bias. The dashed lines show fits of (2.15) to the experimental data. Error bars represent 99 % confidence intervals. (●) W=0.25 μm, L=0.18 μm, $\alpha = 0.071$, (■) W=10.0 μm, L=0.18 μm, $\alpha = 0.101$, (♦) W=1.0 μm, L=1.0 μm, $\alpha = 0.331$, (▲) W=0.25 μm, L=7.2 μm, $\alpha = 0.342$

Based on these two equations, the following empirical model is proposed:

$$\sigma_{\Delta V_T}(V_{BS}) = \sigma_{\Delta V_T}|_{V_{BS}=0} \cdot \left(1 - \frac{V_{BS}}{\phi_B}\right)^{\alpha}, \tag{2.15}$$

where α is a fitting parameter. It follows that in case of dominating doping fluctuations, $\alpha = 1/4$. For dominant fluctuations in oxide capacitance, $\alpha = 1/2$. For short transistors the threshold voltage becomes less sensitive to the bulk bias and α is expected to decrease. To first order, the width and length dependence of α is modeled by:

$$\alpha(W, L) = A_{0,\alpha}^2 + \frac{A_{L,\alpha}}{L} + \frac{A_{W,\alpha}}{W} + \frac{A_{WL,\alpha}}{WL}, \tag{2.16}$$

in which $A_{0,\alpha}^2$, $A_{L,\alpha}$, $A_{W,\alpha}$ and $A_{WL,\alpha}$ are proportionality constants. This equation takes into account possible deviations for short-, narrow-, and short-and-narrow-channel transistors.

Figure 2.8a shows fits of (2.15) to experimentally obtained values of $\sigma_{\Delta V_T}$ as a function of the bulk bias for several pair dimensions. Figure 2.8b shows the correlation of $\Delta V_T(V_{BS})$ with $\Delta V_T@V_{BS} = 0$ V. It is seen that the bulk bias dependence of $\sigma_{\Delta V_T}$ is well described. For long transistor pairs $\alpha \approx 0.3$. This suggests that threshold voltage mismatch is mainly caused by doping fluctuations. The correlation is seen to drop for decreasing bulk bias, which is expected considering the analysis related to figure 2.7. However, in chapter 4 it will be found that the complete picture is more complicated and that we cannot jump to the conclusion of dominating doping fluctuations this easily. It is mainly for this reason that, at this stage, the empirical model (2.15) is used. Physics- and

technology-related details will be discussed later. For the short transistors it is observed that the bulk bias dependence of threshold voltage mismatch disappears, as expected.

2.3.3 Impact of current factor mismatch

To calculate the influence of a mismatch in the current factor on the drain current, a more detailed description of the drain current is needed. As was reasoned in the previous subsection, current factor mismatch is only expected to have an influence in the strong inversion regime. The following widely used strong inversion model for the drain current is chosen:

$$I_D = \beta(V_{GS} - V_T - V_{DS}/2)V_{DS}. \tag{2.17}$$

The current factor itself is given by:

$$\beta = \frac{WC_{ox}\mu(V_{GS}, V_{DS})}{L}. \tag{2.18}$$

Therefore, mismatch in the current factor can be attributed to mismatch in the transistor dimensions, mismatch in the oxide capacitance or mismatch in the mobility μ, which is the only bias dependent quantity in this equation. It will be seen that mismatch in series resistance is taken into account by an apparent mismatch in the mobility. Before applying (2.3) to (2.17), the bias dependence of μ will be examined.
In strong inversion the mobility is determined by the bulk mobility (μ_B), phonon scattering ($\mu_{ph} = a_{ph}/(V_{GS} - V_T - V_{DS}/2)$), surface roughness scattering ($\mu_{sr} = a_{sr}/(V_{GS} - V_T - V_{DS}/2)$) and velocity saturation ($\mu_{sat} = Lv_{sat}/V_{DS}$), where a_{ph} and a_{sr} are proportionality constants and v_{sat} is the saturation velocity. The given bias dependencies should be considered as first order approximations, which are convenient since they will result in a simple mismatch model. A more accurate analysis of mobility determining effects can be found in e.g. [52–58]. The total mobility is calculated by applying Matthiessen's rule:

$$\frac{1}{\mu} = \frac{1}{\mu_B} + \frac{1}{\mu_{ph}} + \frac{1}{\mu_{sr}} + \frac{1}{\mu_{sat}}. \tag{2.19}$$

Combining this with (2.18) yields:

$$\frac{1}{\beta} = \frac{1}{\beta_0} + \frac{V_{GS} - V_T - V_{DS}/2}{\zeta_{sr}} + \frac{V_{DS}}{\zeta_{sat}}, \tag{2.20}$$

where $\beta_0 = WC_{ox}\mu_B/L$, $1/\zeta_{sr} = (L/WC_{ox})((1/a_{ph}) + (1/a_{sr}))$ and $\zeta_{sat} = WC_{ox}v_{sat}$. Mobility depends only weakly on the bulk bias. This

dependence will therefore be neglected. Mathematically, (2.20) is equivalent to the approach followed by [17, 31, 37–40, 18], in which the current factor is described by:

$$\beta = \frac{\beta_0}{1 + \theta_{sr}(V_{GS} - V_T - V_{DS}/2) + \theta_{sat}V_{DS}}, \tag{2.21}$$

where the mobility reduction parameters $\theta_{sr} = \beta_0/\zeta_{sr}$ and $\theta_{sat} = \beta_0/\zeta_{sat}$. The parameters θ_{sr} and θ_{sat} do not depend on the oxide capacitance, but they do depend on μ_B. In our formulation β_0, ζ_{sr} and ζ_{sat} all depend on C_{ox}, but the mobility determining effects are represented by separate parameters.

When series resistance at the source (R_S) and drain (R_D) plays a significant role, in (2.17) and (2.20) V_{GS} needs to be replaced by $V_{GS} - I_D R_S$ and V_{DS} by $V_{DS} - I_D(R_S + R_D)$. Since the MOS transistor is symmetrical, $R_S = R_D$. When the impact of series resistance on the current factor, described by (2.20), is neglected, it easily follows that series resistance effects can be included by replacing the parameters ζ_{sr} and ζ_{sat} by:

$$1/\zeta_{sr} = (L/WC_{ox})((1/a_{ph}) + (1/a_{sr})) + R_S + R_D, \tag{2.22}$$

$$1/\zeta_{sat} = (1/WC_{ox}v_{sat}) + (R_S - R_D)/2. \tag{2.23}$$

It is seen that the impacts of the source and drain resistance on ζ_{sat} cancel out. However, the mismatches in source and drain resistance (ΔR_S and ΔR_D) are uncorrelated and do not have to be equal within one transistor. Therefore, the fluctuations in ζ_{sat} are affected by mismatch in series resistance.

All of the above presented equations are valid in the linear regime of the saturation region, which means that the drain bias is smaller than the saturation voltage (V_{DSsat}). For larger drain bias it has to be replaced by the saturation voltage, which is calculated by putting $dI_D/dV_{DS} = 0$. Applying this to (2.17) with (2.20) yields:

$$V_{DSsat} = \frac{\sqrt{a^2 + 2ab(V_{GS} - V_T)} - a}{b}, \tag{2.24}$$

where $a = (1/\beta_0) + (V_{GS} - V_T)/\zeta_{sr}$ and $b = (1/\zeta_{sat}) - (1/2\zeta_{sr})$. When $b \to 0$, this equation simplifies to:

$$V_{DSsat} = V_{GS} - V_T. \tag{2.25}$$

Even when $b \to 0$ it is preferred to use (2.24), because the use of (2.25) creates a discontinuity between the linear and the saturation regime.

We will now proceed by calculating the impact of a mismatch in the current factor on the drain current. The parameters determining current factor mismatch are the mismatch in β_0, ζ_{sr} and ζ_{sat}. Applying (2.3) to (2.17) with (2.20) yields:

$$\left.\frac{\Delta I_D}{I_D}\right|_{\Delta(1/\beta)} = -\beta\Delta\frac{1}{\beta_0} - \beta(V_{GS} - V_T - V_{DS}/2)\Delta\frac{1}{\zeta_{sr}} - \beta V_{DS}\Delta\frac{1}{\zeta_{sat}}. \tag{2.26}$$

This equation is valid in the linear regime. In saturation, again, V_{DS} needs to be replaced by V_{DSsat}. Fluctuations in V_{DSsat} do not influence the drain current, since $\partial I_D/\partial V_{DSsat} = 0$.

2.3.4 The complete model

The total mismatch in the drain current is calculated by adding the contribution due to threshold voltage-mismatch (2.7) and the contribution due to current-factor mismatch (2.26):

$$\frac{\Delta I_D}{I_D} = \left.\frac{\Delta I_D}{I_D}\right|_{\Delta V_T} + \left.\frac{\Delta I_D}{I_D}\right|_{\Delta(1/\beta)} \tag{2.27}$$

It was found earlier that the first term on the right hand side is approximately valid from $V_{GS} = V_T$ to $V_{GS} = V_{DD}$, while the second term is only valid in strong inversion. However, as was reasoned before, at low gate bias $\Delta I_D/I_D|_{\Delta V_T}$ is much larger than $\Delta I_D/I_D|_{\Delta(1/\beta)}$. It is therefore safe to use (2.26) for all gate biases greater than the threshold voltage without much loss of overall model accuracy. Extrapolating to even lower gate biases results in negative mobility terms, which might lead to singularities. To avoid this, when a mobility term in (2.20) turns negative, it is equated to zero, which is equivalent to removing it from the model.

To calculate $\mu_{\Delta I_D/I_D}$ and $\sigma_{\Delta I_D/I_D}$, (2.4) and (2.5) need to be applied to (2.27). For $\sigma_{\Delta I_D/I_D}$ this results in:

$$\sigma^2_{\Delta I_D/I_D} = \left(\frac{g_m}{I_D}\right)^2 \sigma^2_{\Delta V_T} + \beta^2\sigma^2_{\Delta(1/\beta_0)} + \beta^2(V_{GS} - V_T - V_{DS}/2)^2\sigma^2_{\Delta(1/\zeta_{sr})} + \tag{2.28}$$

$$+ \beta^2 V_{DS}^2\sigma^2_{\Delta(1/\zeta_{sat})} + \text{correlation terms}.$$

The width and length dependence of the variances of the mismatch parameters and the correlation factors will be modeled in the next section. We will now proceed with the development of the parameter extraction routine.

2.3.5 Parameter extraction

In the previous subsections two kinds of parameters were encountered, parameters related to the modeling of the drain current (V_T, β_0, ζ_{sr} and ζ_{sat}) and the parameters describing the mismatch (ΔV_T, $\Delta(1/\beta_0)$, $\Delta(1/\zeta_{sr})$ and $\Delta(1/\zeta_{sat})$). Firstly, the extraction of drain-current-model parameters will be outlined. Secondly we will look into the extraction of the mismatch-model parameters.

Drain-current-model parameters. Of the parameters V_T, β_0, ζ_{sr} and ζ_{sat}, the first three are estimated from the $I_D - V_{GS}$ curve in the linear regime at low drain bias ($V_{DS} = 50$ mV) at which the ζ_{sat} term in (2.20) can be ignored. First ζ_{sr} is considered to be infinite and V_T and β_0 are determined by the maximum slope method[6]:

$$\beta_0 = \frac{1}{1 - I_D/(\zeta_{sr} V_{DS})} \cdot \frac{g_m}{V_{DS}} \;@\; g_m = g_{mmax} \tag{2.29}$$

$$V_T = V_{GS} - V_{DS}/2 - \frac{I_D}{\beta_0 V_{DS}} \;@\; g_m = g_{mmax}, \tag{2.30}$$

where g_{mmax} is the maximum transconductance. The first factor on the right hand side of (2.29) is a correction for finite values of ζ_{sr}, which lower the transconductance. To estimate ζ_{sr}, (2.17) with (2.20) is rewritten into:

$$I_D(1+(\beta_0/\zeta_{sr})(V_{GS}-V_T-V_{DS}/2)) = \beta_0(V_{GS}-V_T-V_{DS}/2)V_{DS}, \tag{2.31}$$

from which $1/\zeta_{sr}$ is estimated by a linear least squares fit. Since this model is only valid in the strong inversion region, the fit ranges from $V_{GS}@g_m = g_{mmax}$ to $V_{GS} = V_{DD}$. This new value of ζ_{sr} is now introduced into (2.29) after which new values of β_0 and V_T are calculated. From these values a new value of ζ_{sr} can be calculated and so on. This process is iterated until no significant changes are observed.
The parameter ζ_{sat} is extracted from the $I_D - V_{GS}$ curve in saturation ($V_{DS} = V_{DD}$). The approach is the same as the one followed to extract ζ_{sr}. Multiplying the left and right hand sides of (2.17) by $1/\beta$ results in a function that depends on $1/\zeta_{sat}$, like (2.31) depends on $1/\zeta_{sr}$. From this, $1/\zeta_{sat}$ is extracted by a least squares fit. Since the estimation takes place in strong inversion, in (2.17) and (2.20) V_{DS} needs to be replaced by (V_{DSsat}), which is itself a function of ζ_{sat}. To avoid this problem the

[6]The maximum slope method will be illustrated in chapter 3, figure 3.1.

Table 2.2. Extracted values of V_T, β_0, ζ_{sr} and ζ_{sat} for several device dimensions. The oxide thickness is equal to 2.8 nm.

W (μm)	L (μm)	V_T (V)	β_0 (μA V^{-2})	ζ_{sr} (μA V^{-1})	ζ_{sat} (μA V^{-1})
0.25	0.18	0.282	516	$1.70 \cdot 10^3$	996
10.0	0.18	0.329	$17.8 \cdot 10^3$	$45.3 \cdot 10^3$	$36.0 \cdot 10^3$
1.0	1.0	0.373	279	$3.17 \cdot 10^3$	$2.75 \cdot 10^3$
0.25	7.2	0.293	10.6	156	136

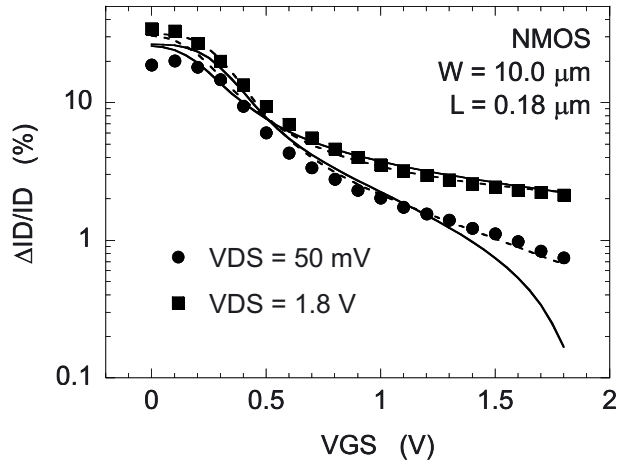


Figure 2.9. The $\Delta I_D/I_D - V_{GS}$ curves of a certain transistor pair at $V_{DS} = 50$ mV and $V_{DS} = V_{DD}$. Also shown is a standard least squares fit to these curves (full lines) and a weighted least squares fit (dashed lines). The drain-current-model parameters are listed in table 2.2.

following estimate for the saturation voltage is used:

$$V_{DSsat} \cong \sqrt{2 \left.\frac{(V_{GS} - V_T - V_{DS}/2)V_{DS}}{I_D}\right|_{V_{DS}=50 \text{ mV}} \cdot I_D|_{V_{DS}=V_{DD}}}. \tag{2.32}$$

The extracted parameters are summarized in table 2.2.

Mismatch-model parameters. We will proceed with the extraction of ΔV_T, $\Delta(1/\beta_0)$, $\Delta(1/\zeta_{sr})$ and $\Delta(1/\zeta_{sat})$. It follows from (2.7), (2.26) and (2.27) that $\Delta I_D/I_D$ depends linearly on these parameters. The most straightforward way to extract the parameters is to use a linear fit of these equations to the experimental $\Delta I_D/I_D - V_{GS}$ curves. Both the

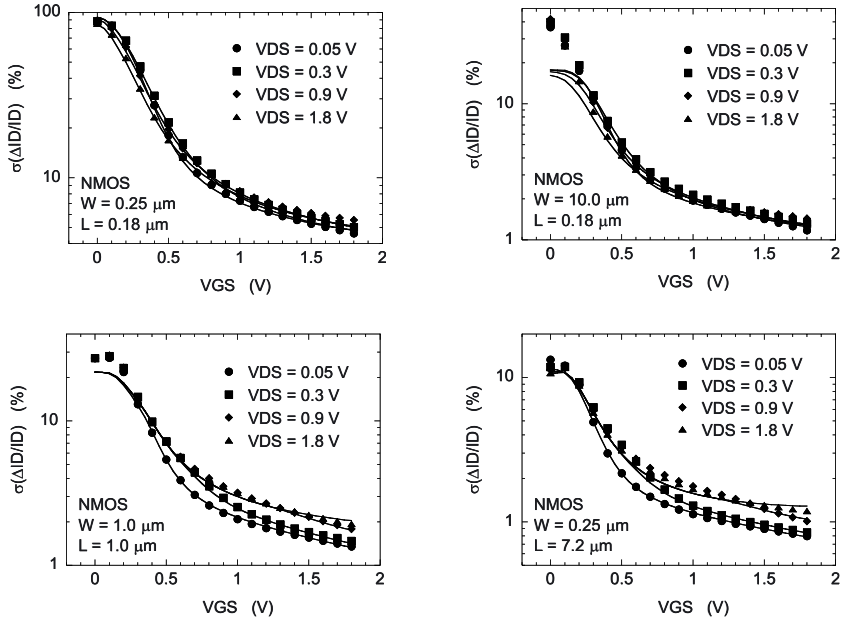
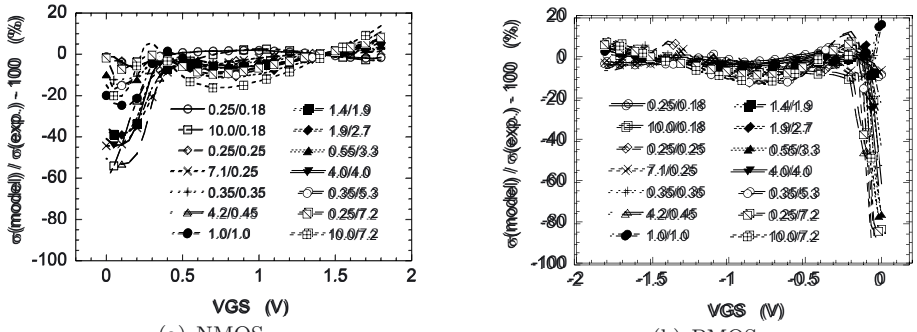


Figure 2.10. Experimental (symbols) and modeled (lines) $\sigma_{\Delta I_D/I_D} - V_{GS}$ curves for several values of the drain bias and device dimensions. $V_{BS} = 0$ V



(a) NMOS (b) PMOS

Figure 2.11. Relative difference between the modeled and experimental $\sigma_{\Delta I_D/I_D} - V_{GS}$ curves for all measured device dimensions. The W/L ratios are included in the plots. $|V_{DS}| = 1.8$ V, $V_{BS} = 0$ V

curve at $V_{DS} = 50$ mV and at $V_{DS} = V_{DD}$ are included in the fit. The gate bias ranges from the minimum out of figure 2.5a to $V_{GS} = V_{DD}$. A disadvantage of this method is illustrated in figure 2.9, which shows the mismatch in one transistor pair as a function of the gate bias. It is observed that the mismatch at low gate bias is much higher than

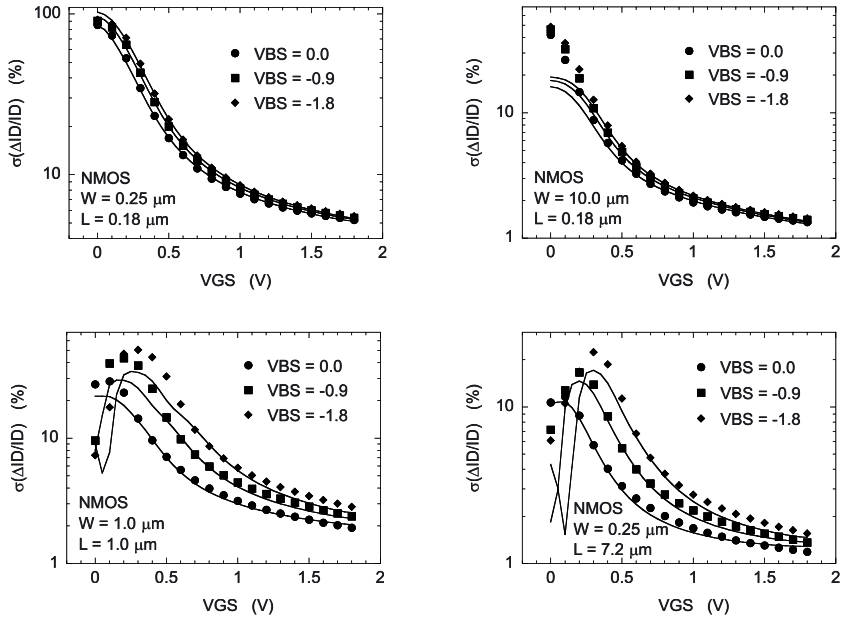
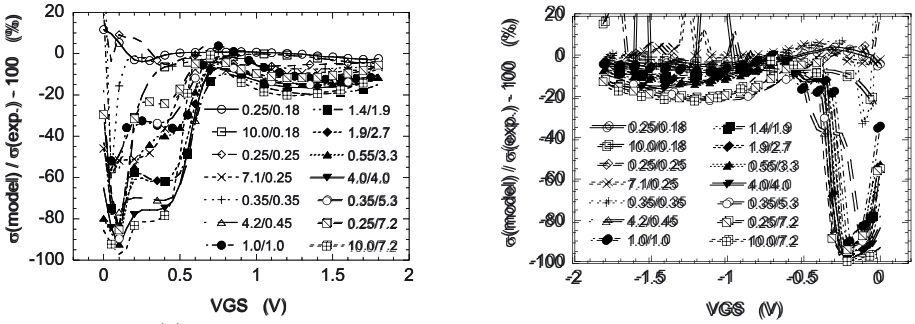


Figure 2.12. Experimental (symbols) and modeled (lines) $\sigma_{\Delta I_D/I_D} - V_{GS}$ curves for several values of the bulk bias and device dimensions. $V_{DS} = 1.8$ V



(a) NMOS (b) PMOS

Figure 2.13. Relative difference between the modeled and experimental $\sigma_{\Delta I_D/I_D} - V_{GS}$ curves for all measured device dimensions. The W/L ratios are included in the plots. $|V_{DS}| = |-V_{BS}| = 1.8$ V

at high gate bias. Therefore, a small inaccuracy in the model at low gate bias could have a large impact on the obtained relative accuracy at higher gate biases. To avoid this problem, we choose not to minimize the sum of the squared differences, but to minimize the sum of the squared relative differences. This normalization is achieved by dividing the left-

and right-hand sides of (2.27) by $\sigma_{\Delta I_D/I_D}$ before performing the fit. The result of this fit is also shown the figure 2.9.

2.3.6 Model accuracy

This subsection examines the accuracy of the developed model by comparing experimental $\sigma_{\Delta I_D/I_D} - V_{GS}$ curves to the modeled ones. Figure 2.10 shows the comparison at zero bulk bias with the drain bias as a parameter. For all shown dimensions it is seen that the curves at $V_{DS} = 50$ mV and $V_{DS} = 1.8$ V are well described by the model. The curves at $V_{DS} = 0.3$ V and $V_{DS} = 0.9$ V were not included in the fit and are seen to be well predicted. Figure 2.11 shows the model accuracy for all measured dimensions at zero bulk bias and $|V_{DS}| = 1.8$ V. In the strong inversion region the model is seen to describe the measurements within the required 20 % accuracy. In weak inversion the accuracy decreases as expected (see subsection 2.3.2 and figure 2.5).
Figure 2.12 compares experimental and modeled $\sigma_{\Delta I_D/I_D} - V_{GS}$ curves at $V_{DS} = 1.8$ V with the bulk bias as a parameter. Again the curves are seen to be well described. Note that all of the parameters, except α, have been extracted at zero bulk bias. The values for α are taken from figure 2.8a. Figure 2.13 shows the model accuracy for all measured pair dimensions at $|V_{DS}| = |-V_{BS}| = 1.8$ V. Again, within the strong inversion region the model is seen to describe the measurements within the required 20 % accuracy range. In weak inversion the accuracy decreases. Note that in this figure the weak inversion region is larger than that in figure 2.11, since the threshold voltage increases with decreasing bulk bias.

2.4 Width and length dependence

In the previous section a model was developed to describe the mismatch in the drain currents of a transistor pair. In this section the width and length dependence will be modeled of the variance of the mismatch in a certain parameter ($\sigma^2_{\Delta P}$) (subsection 2.4.1) and of the correlation factors between the mismatches in parameters (subsection 2.4.2). The last subsection tests and demonstrates the derived model.

2.4.1 Width and length dependence of $\sigma^2_{\Delta P}$

Several publications exist that deal with modeling $\sigma^2_{\Delta P}(W, L)$. The most referred to is the one published by Pelgrom *et al.* [5]. This subsection is started by presenting a summary of this work.
It is assumed that the parameter P can locally be defined as $P(x, z) = \mu_P + \delta P(x, z)$ and that the overall transistor parameter P is given by averaging $P(x, z)$ over the area of the transistor. Now assume

that the first device is located between the coordinates $\{x_1, z_1\}$ and $\{x_1 + L, z_1 + W\}$ and that the second device is located between $\{x_2, z_2\}$ and $\{x_2 + L, z_2 + W\}$. The mismatch between the two devices is then given by:

$$\Delta P = \frac{1}{WL}\left(\iint_{\{x_2,z_2\}}^{\{x_2+L,z_2+W\}} \delta P(x',z')dx'dz' - ... \right. \tag{2.33}$$

$$\left. ... \iint_{\{x_1,z_1\}}^{\{x_1+L,z_1+W\}} \delta P(x',z')dx'dz' \right).$$

This equation can be interpreted as the convolution of a mismatch causing disturbance function $P(x, z)$ and a geometry function $G(x, z)$, which is given by:

$$G(x,z) = \begin{cases} \frac{-1}{WL} & \{x,z\} \in \{\{x_1,z_1\},\{x_1+L,z_1+W\}\} \\ \frac{1}{WL} & \{x,z\} \in \{\{x_2,z_2\},\{x_2+L,z_2+W\}\} \quad i = 1,2. \\ 0 & \{x,z\} \notin \{\{x_i,z_i\},\{x_i+L,z_i+W\}\} \end{cases} \tag{2.34}$$

Convolution in the space domain is equivalent to multiplication in the spacial frequency domain:

$$\Delta\mathcal{P}(\omega_x,\omega_z) = \mathcal{G}(\omega_x,\omega_z)\cdot\delta\mathcal{P}(\omega_x,\omega_z), \tag{2.35}$$

where $\mathcal{G}(\omega_x, \omega_z)$ and $\delta\mathcal{P}(\omega_x, \omega_z)$ are the two-dimensional Fourier transforms of $G(x, z)$ and $\delta P(x, z)$, respectively. From (2.34) the first is calculated to be:

$$\mathcal{G}(\omega_x,\omega_z) = \frac{\sin(L\omega_x/2)\sin(W\omega_z/2)}{(L\omega_x/2)(W\omega_z/2)}\left(e^{i(x_2+L/2)\omega_x+i(z_2+W/2)\omega_z} - \dots \right. \tag{2.36}$$

$$\left. \dots e^{i(x_1+L/2)\omega_x+i(z_1+W/2)\omega_z}\right)$$

From basic spectral theory it follows that $\sigma^2_{\Delta P}$ is equal to:

$$\sigma^2_{\Delta P} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\Delta\mathcal{P}(\omega_x,\omega_z)|^2 d\omega_x d\omega_z = \tag{2.37}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\mathcal{G}(\omega_x,\omega_z)|^2\cdot|\delta\mathcal{P}(\omega_x,\omega_z)|^2 d\omega_x d\omega_z,$$

where $|\delta\mathcal{P}(\omega_x, \omega_z)|^2$ is the power spectrum of $\delta P(x, z)$. When the lowest significant frequency of the mismatch generating process is much larger

than $1/W$ and $1/L$, a mismatch causing event in one device does not have an impact on the other device and (2.37) can be approximated by:

$$\sigma^2_{\Delta P} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\mathcal{G}(\omega_x,\omega_z)|^2 \cdot |\delta\mathcal{P}(0,0)|^2 d\omega_x d\omega_z = \tag{2.38}$$

$$= \frac{8\pi^2|\delta\mathcal{P}(0,0)|^2}{WL} \equiv \frac{A^2_{0,\Delta P}}{WL}.$$

Summarizing, it is found that $\sigma^2_{\Delta P}$ is proportional to $1/WL$ and the proportionality constant $A^2_{0,\Delta P}$.
Until now we have implicitly assumed that the parameter under consideration is itself not a function of the width and length. If this is not the case, the model needs to be adapted. We will assume that the width and length dependence of the parameter $P(x,z)$ can be written as:

$$P(W,L,x,z) = \mathrm{f}(W,L) \cdot P'(x,z), \tag{2.39}$$

where $\mathrm{f}(W,L)$ models the width and length dependence of P, but is constant in space, and $P'(x,z)$ is independent of the width or length, but does vary with x and z. It easily follows that in this case:

$$\sigma^2_{\Delta P} = \mathrm{f}(W,L)^2 \frac{A^2_{0,\Delta P'}}{WL}, \tag{2.40}$$

The equations presented until now are valid for long and wide transistors. Corrections are required for short and narrow transistors [16, 17, 19, 22, 29–34, 37, 38, 40, 51, 59–65]. The physical aspects of these short- and narrow-channel effects will be discussed in chapter 4. Here, we will limit ourselves to the following empirical description:

$$\sigma^2_{\Delta P} = \frac{A^2_{0,\Delta P}}{WL} + \frac{A_{L,\Delta P}}{WL^2} + \frac{A_{W,\Delta P}}{W^2L} + \frac{A_{WL,\Delta P}}{W^2L^2}, \tag{2.41}$$

where the terms containing $A_{L,\Delta P}$, $A_{W,\Delta P}$ and $A_{WL,\Delta P}$ describe to first order deviations for short, narrow and short and narrow device pairs, respectively. When P is a function of the width or length, the right hand side again needs to be multiplied by $\mathrm{f}(W,L)^2$.

2.4.2 Width and length dependence of correlation factors

When the mismatches in two parameters are correlated, this means that they are partially determined by the same mismatch causing mechanism. For long and wide transistors the mechanisms determining the operation of the MOS transistor do not vary with width and length.

Therefore, the correlation factor is expected to be constant. Again, for short or narrow transistors deviations can be expected, which are modeled as follows:

$$\rho(\Delta P_1, \Delta P_2) = A^2_{0,\Delta P_1,\Delta P_2} + \frac{A_{L,\Delta P_1,\Delta P_2}}{L} + \frac{A_{W,\Delta P_1,\Delta P_2}}{W} + \frac{A_{WL,\Delta P_1,\Delta P_2}}{WL}. \tag{2.42}$$

Note that the correlation factor, and therefore $A^2_{0,\Delta P_1,\Delta P_2}$, can be negative.

2.4.3 Matching properties of a 0.18 μm CMOS process

In this subsection the theory developed in the previous two subsections will be applied to the 0.18 μm CMOS technology, that was briefly introduced in section 2.2. The variances of the mismatch-model parameters (ΔV_T, $\Delta(1/\beta_0)$, $\Delta(1/\zeta_{sr})$ and $\Delta(1/\zeta_{sat})$), the correlation between these parameters and the parameter α, related to the bulk bias dependence of $\sigma_{\Delta V_T}$, were already extracted in section 2.3. The parameters A^2_0, A_L, A_W and A_{WL} related to the variances are extracted by a linear weighted least squares fit of (2.41) to the experimentally obtained $\sigma^2_{\Delta P}$'s for different widths and lengths. The weight attributed to each point is equal to $1/\sigma^2(\sigma^2_{\Delta P})$, which can be calculated using (2.1). Since the current factor is proportional to W/L, the parameters $\sigma^2_{\Delta(1/\beta_0)}$, $\sigma^2_{\Delta(1/\zeta_{sr})}$ and $\sigma^2_{\Delta(1/\zeta_{sat})}$ are multiplied by $(W/L)^2$ prior to the fit. The parameters related to the width and length dependence of the correlation factors and α are extracted by a normal linear least squares fit of (2.42) to the experimental data.

Besides extracting A^2_0, A_L, A_W and A_{WL}, standard regression analysis is applied to determine the standard deviations of these parameters. When this standard deviation is larger than the absolute value of the parameter itself, it is concluded that the obtained value is not significant. In such a case the parameter is removed from the model by equating it to zero after which the fit and regression analysis are repeated.

The results of this exercise are presented in table 2.3. Figure 2.14 compares the experimental and modeled width and length dependence of $\sigma_{\Delta V_T}$, $\sigma_{\Delta(1/\beta_0)}$, $\sigma_{\Delta(1/\zeta_{sr})}$ and $\sigma_{\Delta(1/\zeta_{sat})}$ and figure 2.15 shows the comparison for the most significant correlation factors. It is observed that (2.41) and (2.42) provide a good description of the experimental data. It follows from table 2.3 that the main part of the mismatch in the drain current is caused by $\sigma_{\Delta V_T}$ and $\sigma_{\Delta(1/\beta_0)}$. We will briefly discuss the obtained results. A more thorough investigation of the physical and technological origins of MOSFET mismatch will be presented in chapter

Table 2.3. Extracted parameters for describing the width and length dependence of the variances of the mismatch-model parameters and the correlation coefficients between these parameters. 1σ confidence intervals are included. Dimensions are such that the width and length are given in μm, $\sigma_{\Delta V_T}$ in mV, $\sigma_{\Delta(1/\beta_0)}$ in ΩV and $\sigma_{\Delta(1/\zeta_{sr})}$ and $\sigma_{\Delta(1/\zeta_{sat})}$ in Ω. Transistor-model parameters are listed in table 2.2.

	A_0^2	A_L	A_W	A_{WL}
		NMOS		
$\sigma^2_{\Delta V_T}$	36.6 ± 2.8	4.87 ± 0.97	-4.0 ± 1.0	0
$\sigma^2_{\Delta(1/\beta_0)}$	$4.92e3 \pm 0.75e3$	$0.93e3 \pm 0.42e3$	$-0.47e3 \pm 0.29e3$	$-0.29e3 \pm 0.12e3$
$\sigma^2_{\Delta(1/\zeta_{sr})}$	$0.92e3 \pm 0.11e3$	61 ± 49	-141 ± 39	-19.1 ± 13.9
$\sigma^2_{\Delta(1/\zeta_{sat})}$	880 ± 97	0	-136 ± 32	0
$\rho_{\Delta V_T,\Delta(1/\beta_0)}$	0.045 ± 0.036	-0.016 ± 0.013	0	0
$\rho_{\Delta V_T,\Delta(1/\zeta_{sr})}$	-0.186 ± 0.033	0.050 ± 0.015	0	-0.0182 ± 0.0045
$\rho_{\Delta V_T,\Delta(1/\zeta_{sat})}$	-0.587 ± 0.046	0.042 ± 0.017	0.093 ± 0.023	-0.0099 ± 0.0067
$\rho_{\Delta(1/\beta_0),\Delta(1/\zeta_{sr})}$	-0.923 ± 0.023	0.0488 ± 0.0071	0.0169 ± 0.0092	0
$\rho_{\Delta(1/\beta_0),\Delta(1/\zeta_{sat})}$	0	0	0	0
$\rho_{\Delta(1/\zeta_{sr}),\Delta(1/\zeta_{sat})}$	0.049 ± 0.036	-0.065 ± 0.017	0	0.0098 ± 0.0050
α	0.338 ± 0.041	-0.030 ± 0.015	0.054 ± 0.021	-0.0162 ± 0.0061
		PMOS		
$\sigma^2_{\Delta V_T}$	11.8 ± 1.2	6.1 ± 1.3	0	-0.89 ± 0.39
$\sigma^2_{\Delta(1/\beta_0)}$	$43.5e3 \pm 4.0e3$	$-3.64e3 \pm 0.87e3$	$-2.9e3 \pm 1.1e3$	0
$\sigma^2_{\Delta(1/\zeta_{sr})}$	$8.7e3 \pm 1.5e3$	$-0.63e3 \pm 0.35e3$	$-1.09e3 \pm 0.38$	0
$\sigma^2_{\Delta(1/\zeta_{sat})}$	$3.30e3 \pm 0.55e3$	$0.85e3 \pm 0.46e3$	0	$-0.21e3 \pm 0.12e3$
$\rho_{\Delta V_T,\Delta(1/\beta_0)}$	-0.161 ± 0.055	-0.077 ± 0.025	0	0.0109 ± 0.0076
$\rho_{\Delta V_T,\Delta(1/\zeta_{sr})}$	0.191 ± 0.039	-0.172 ± 0.018	0	0.0138 ± 0.0054
$\rho_{\Delta V_T,\Delta(1/\zeta_{sat})}$	-0.173 ± 0.040	0.155 ± 0.019	0	-0.0121 ± 0.0056
$\rho_{\Delta(1/\beta_0),\Delta(1/\zeta_{sr})}$	-0.66 ± 0.11	0.281 ± 0.040	0	0
$\rho_{\Delta(1/\beta_0),\Delta(1/\zeta_{sat})}$	-0.123 ± 0.098	-0.198 ± 0.035	0	0
$\rho_{\Delta(1/\zeta_{sr}),\Delta(1/\zeta_{sat})}$	0	-0.143 ± 0.036	0	0
α	0.294 ± 0.042	-0.046 ± 0.015	0	0

4 and 5, respectively.

Firstly note that the value of $A_{0,\Delta V_T}$ for the NMOS transistors is roughly 50 % higher than those published in literature for 0.18 μm CMOS technologies [6]. This is mainly due to the choice of an amorphous silicon gate material instead of using a fine-grain poly-silicon gate [21, 66]. This

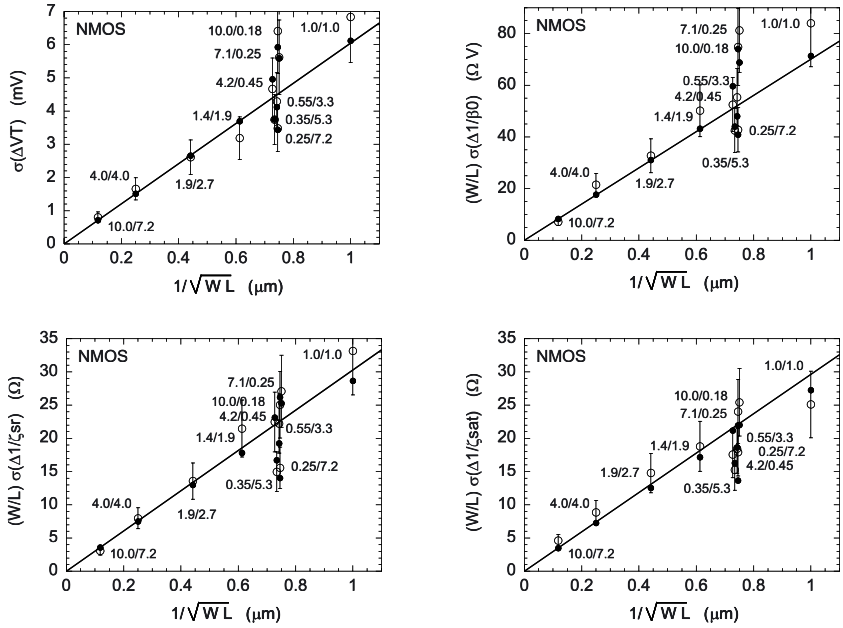


Figure 2.14. Experimental (open symbols) and modeled (full symbols) values of $\sigma_{\Delta V_T}$, $\sigma_{\Delta(1/\beta_0)}$, $\sigma_{\Delta(1/\zeta_{sr})}$ and $\sigma_{\Delta(1/\zeta_{sat})}$ as a function of $1/\sqrt{WL}$. n the plots the W/L ratios of the transistors are given in μm/μm. The full line represents the modeled result if only A_0 is taken into account. Error bars represent 99 % confidence intervals.

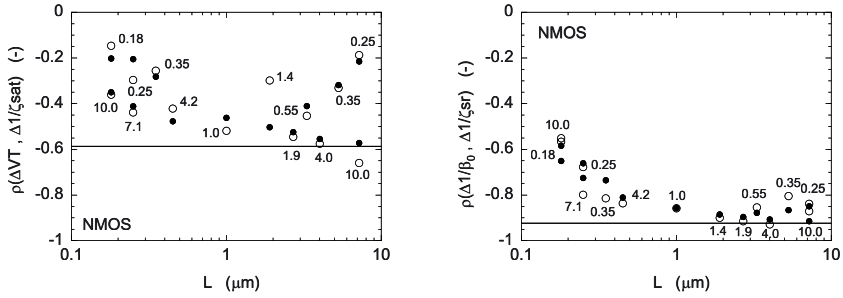


Figure 2.15. Experimental (open symbols) and modeled (full symbols) values of $\rho(\Delta V_T, \Delta(1/\zeta_{sat}))$ and $\rho(\Delta(1/\beta_0), \Delta(1/\zeta_{sr}))$ as a function of the length. In the plots the width of the transistors is given in μm. The full line represents the modeled result if only A_0 is taken into account.

will be experimentally verified in section 5.2. The reasonably low value of $A_{0,\Delta V_T}$ for PMOS devices indicates that the boron gate doping at the gate-oxide interface is uniform and high enough for gate-depletion effects to be under control. The nitrided gate oxide effectively prevents boron

penetration, which would seriously degrade the matching performance. The relatively large increase in mismatch for short-channel PMOS transistors is due to the absence of halos, which causes the effective channel length to be significantly smaller than the metallurgical channel length. Secondly, it is observed that the mismatch decreases for narrow NMOS transistors. This could be caused by a commonly observed lower boron doping concentration close to the shallow-trench isolation. Generally, this narrow-channel effect is less pronounced for PMOS transistors. Therefore, the decreases in mismatch are less significant or not significant at all.
Thirdly, significant negative correlations are observed between ΔV_T and $\Delta(1/\zeta_{sat})$ and between $\Delta(1/\beta_0)$ and $\Delta(1/\zeta_{sr})$, which do not have a clear physical origin. We therefore conclude that they are mainly related to inaccuracies caused by the simplicity of the model. The correlation between $\Delta(1/\beta_0)$ and $\Delta(1/\zeta_{sr})$ is caused by the simplified expressions for the mobility. More complicated expressions will be presented in chapter 4. The correlation between ΔV_T and $\Delta(1/\zeta_{sat})$ is caused by neglecting the dependence of the local threshold voltage on its lateral position. Again, this will be further discussed in chapter 4.

2.5 Example: Yield of a current-steering D/A converter

A typical circuit that suffers from MOSFET mismatch is the current-steering D/A converter (DAC[7]). In case of a binary implementation of the DAC, the least significant bit consists of a single unit current cell that produces a current (I_{cs}). The most significant bit consists of $2^{N_{BIT}-1}$ unit current cells. It is clear that the total accuracy of the more significant bits should be better than half the least significant bit. This accuracy is determined by the device dimensions of the unit current cell. As an illustration of the developed model, the minimum device dimensions of the unit current cell will be calculated for which a certain yield of a DAC is obtained. The approach is based on the one presented in [67]. Firstly, the minimum accuracy of the unit current cell is calculated based on the required number of bits and yield. This minimum accuracy puts a first constraint on the device dimensions. The implications of this constraint can be calculated with the model developed earlier in this chapter. A second constraint is put forward by the required current

[7]In the remainder of this chapter the abbreviation DAC will be used for the current-steering D/A converter.

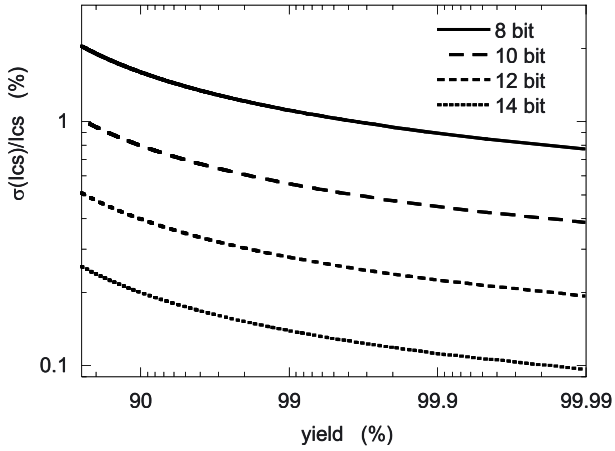


Figure 2.16. Required accuracy of the unit current cell ($\sigma_{I_{sc}}/I_{sc}$) as a function of the yield for an 8, 10, 12, and 14 bits DAC

range at the output of the circuit. Both constraints together determine the dimensions of the unit current cell.

2.5.1 Accuracy of unit current cell based on a yield requirement

The yield of a DAC is related to its integral non-linearity error (INL). This error is defined as the maximum deviation of the output current from the expected current. For the DAC to work properly, the INL should be smaller than $I_{cs}/2$. This results in the following requirement for the relative accuracy of the unit current cell ($\sigma_{I_{cs}}/I_{cs}$) [67]:

$$\frac{\sigma_{I_{cs}}}{I_{cs}} = \frac{1}{\sqrt{2^{N_{BIT}+3}} \cdot \text{InverseErf}\left(\frac{1+yield}{2}\right)}. \tag{2.43}$$

InverseErf is the inverse of the error function.
Figure 2.16 plots the required accuracy as a function of the yield for an 8, 10, 12, and 14 bits DAC. It is observed that to obtain a 99.7 % yield, relative accuracies of approximately 1.0 %, 0.5 %, 0.25 %, and 0.125 % are required, respectively.

2.5.2 Width and length of the unit current cell

The required accuracy of the unit current cell puts a constraint on the dimensions of the current cell. This constraint can be calculated using the model that was derived earlier in this chapter. However, the bias conditions of the transistor that provides the current need to be known.

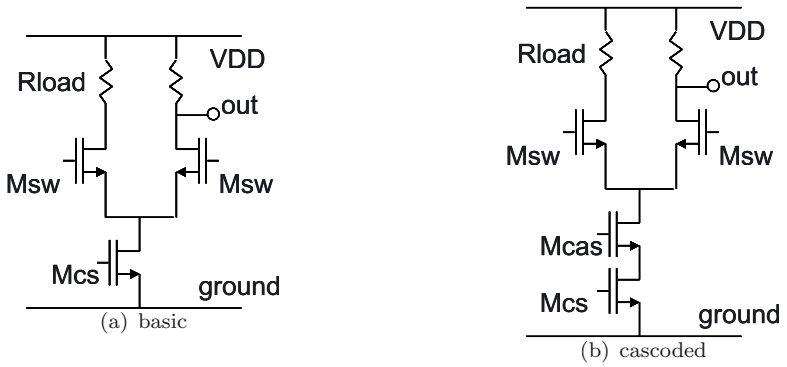


Figure 2.17. Two possible configurations of the unit current cell of a DAC

These depend on the configuration of the unit current cell. Figure 2.17 shows two possible configurations. The first, shown in figure 2.17a, is the basic configuration. This consists of a current providing MOSFET (Mcs) and two MOSFETs (Msw) that make up a switch that directs the current either to the output or to another branch. The second configuration, presented in figure 2.17b, is similar to the first, but Mcs is cascoded by Mcas. This has the advantage of increasing the output impedance and it also increases the maximum frequency of operation. This configuration has the disadvantage that it leaves less head room for Mcs, which determines the accuracy of the unit current cell. It follows from (2.28) and from figure 2.10 that the gate bias of Mcs should be chosen as large as possible to maximize the accuracy. The gate bias is limited by the fact that all transistors in the cell should operate in the saturation regime. For the calculations presented here a gate bias of $V_{GS} = 0.9$ V will be assumed. This results in a gate overdrive of $V_{GS} - V_T = 0.5 - 0.6$ V, depending on the gate length.
Furthermore, the calculations will be based on the parameters presented in table 2.3 that describe the width and length dependence of the variability. Besides these parameters, also the width and length dependencies of g_m/I_D, the threshold voltage, and of the current factors need to be know. For the technology under study these can be approximated by the following polynomials[8]:

$$\left.\frac{g_m}{I_D}\right|_{V_{GS}=0.9V} = 3.541 + \frac{0.129}{L} - \frac{0.0758}{L^2} - \frac{0.1096}{W} - \frac{0.0420}{WL} + \frac{0.0107}{WL^2} \text{ V}^{-1} \tag{2.44}$$

[8]In these equations the width and length are in micrometers.

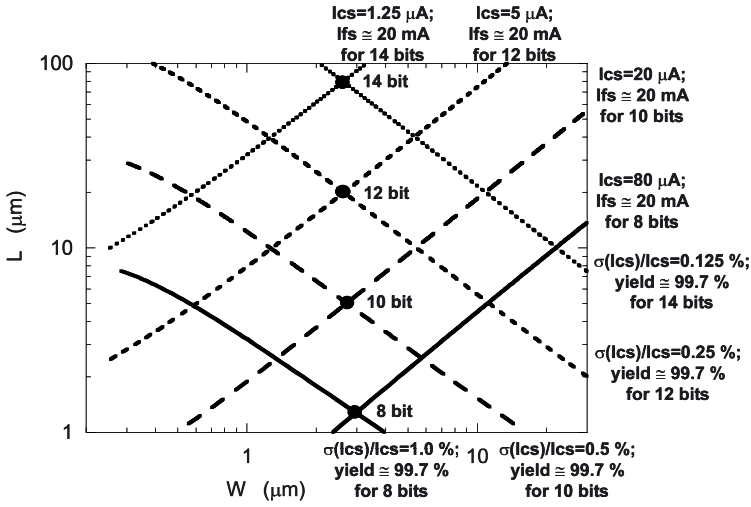


Figure 2.18. Required device length of Msc as a function of the device width for certain constraints. The lines with negative slope give the width-length combinations at which a certain accuracy is reached. The lines with positive slope give the width-length combinations at which a certain current level is reached. The intersection of the relevant lines determines the correct device width and length. For the 8, 10, 12, and 14 bit DAC these intersections are highlighted for the constraints of 99.7 % yield and a full-scale current of $I_{fs} = 20$ mA.

$$V_T = 0.3670 + \frac{0.041}{L} - \frac{0.00848}{L^2} - \frac{0.0204}{W} - \frac{0.0022}{WL} + \frac{0.00067}{WL^2} \text{ V} \quad (2.45)$$

$$\beta_0 = \left(\frac{W}{L}\right) \cdot \left(290.1 + \frac{2.7}{L} + \frac{3.11}{WL}\right) \ \mu\text{AV}^{-2} \quad (2.46)$$

$$\zeta_{sr} = \left(\frac{W}{L}\right) \cdot \left(3712 - \frac{632}{L} + \frac{196}{W}\right) \ \mu\text{AV}^{-1} \quad (2.47)$$

$$\zeta_{sat} = \left(\frac{W}{L}\right) \cdot \left(3298 - \frac{580}{L} + \frac{75}{W}\right) \ \mu\text{AV}^{-1} \quad (2.48)$$

By combining these equations with (2.28), (2.41), and (2.42), one can numerically determine the width-length combinations that result in a certain variability. The result of this exercise is presented in figure 2.18 in which the lines going from upper-left to lower-right give the required length for reaching a certain accuracy at a given width. Note that for not too short or narrow transistors these lines have a slope of -1 on a log-log scale, which means that the area is constant. This conclusion is easily explained by the analysis presented in section 2.4. The in figure

2.18 displayed levels of accuracy correspond with obtaining a 99.7 % yield on an 8, 10, 12, and 14 bits DAC.
Besides the accuracy, the dimensions of Mcs are also limited by the desired full-scale current ($I_{fs} = (2^{N_{BIT}} - 1)I_{cs}$) of the DAC. For the technology under study, the width and length dependence of I_{cs} can be approximated by the following polynomial:

$$I_{cs}|_{V_{GS}=0.9V} = \left(\frac{W}{L}\right)\cdot\left(37.01 - \frac{6.2}{L} + \frac{1.49}{L^2} + \frac{3.17}{W} + \frac{1.24}{WL} - \frac{0.206}{WL^2}\right)\ \mu\text{A}. \tag{2.49}$$

From this one can determine the width-length combinations that result in a certain current level. The result of this exercise is also presented in figure 2.18. The lines going from lower-left to upper-right give the required length to reach a certain current at a given width. For not too short or narrow transistors these lines have a slope of +1 on a log-log scale, which means that the width-to-length ratio is constant. This conclusion follows directly from (2.49). The in figure 2.18 displayed current levels correspond with obtaining a full-scale current of $I_{fs} =$ 20 mA on an 8, 10, 12, and 14 bits DAC.
The width and length of Msc needs to be chosen in such a way that both the accuracy constraint and the current-level constraint are fulfilled. In figure 2.18 this happens where the relevant lines intersect. In the figure these points are highlighted for the 99.7 % yield and $I_{fc} = 20$ mA cases.

2.6 Conclusions

This chapter dealt with two subjects, the measurement of mismatch in the drain current and the modeling of this mismatch. Firstly, our mismatch measurement setup was described and an overview was presented of commonly used test-structures for qualifying the matching properties of a certain technology. Secondly the drain-current-mismatch model was developed. The model was tested on a 0.18 μm CMOS technology.
In the derivation of the model we strived for a physics based one, valid over a large bias range, continuous between different regions of operation and as simple as possible. The relative difference between model and measurement data should be smaller than 20 %. The mismatch in the drain current was assumed to be split up in a contribution due to threshold-voltage mismatch and a contribution due to current-factor mismatch, which were dealt with separately.
In calculating the impact of a mismatch in threshold voltage on the drain current, the only assumption made is that the drain current is a function of the gate-overdrive voltage, but not of the gate bias nor the threshold voltage separately. The resulting model was found to be valid in strong

inversion and in the upper part of the moderate inversion region. In weak inversion deviations were observed. Threshold voltage mismatch was found to depend only weakly on the drain bias. Therefore, this bias dependency was not taken into account. The bulk bias dependence of the mismatch in the threshold voltage is not modeled directly, since this is considered to be non-physical. Instead we modeled the bulk bias dependence of the standard deviation of the mismatch in the threshold voltage. This was done in a semi-empirical way.
The impact on the drain current of a mismatch in the current factor was split up in three contributions, related to mobility limiting effects, i.e. bulk mobility, phonon scattering, surface roughness scattering and velocity saturation. Mismatch in series resistance is also accounted for by the parameters, related to the current factor. The expression for the saturation voltage was properly derived, which resulted in continuity from the linear to the saturation regime.
To extract the mismatch in the model parameters, a weighted least squares fit was introduced as opposed to a normal least squares fit. In this way small modeling errors at low gate biases were prevented from seriously degrading the accuracy at higher values of the gate bias. In strong inversion the model reached the accuracy requirement for all examined drain- and bulk-bias conditions. To obtain good accuracy in the weak inversion region a modeling effort is still required.
The width and length dependence of the extracted variances was described by the model of Pelgrom *et al.*. The correlation factors were reasoned to be independent of width and length. The models were extended to take short and narrow channel effects into account, which leads to an accurate description of the experimental data.

Chapter 3

PARAMETER EXTRACTION

In the previous chapter a model was derived for describing mismatch in the drain current. The presented technique to extract the model parameters was developed in such a way as to give the most accurate description of the mismatch in the drain current. However, does this approach also yield the most meaningful values of the model parameters themselves? The question we have to ask ourselves is whether this is important. The answer depends on the person who is asking. For a circuit designer, the answer is no, since his main goal is *model accuracy*. However, when doing process development, one is more interested in the *physical meaningfulness* of extracted parameters, since it would help to better understand what is happening inside the devices. For process monitoring, the required *measurement time* plays an important role, which is related to *measurement accuracy*. These requirements – model accuracy, physical meaningfulness of parameters and measurement accuracy and time – can, to a certain degree, be in conflict with each other.

In literature, several techniques have been presented, that extract mismatch model parameters [5, 17–19, 28–31, 33–40, 68–70]. They all claim to extract mismatch in threshold voltage and mismatch in current factor, which are considered to be, more or less, well defined physical parameters of the MOS transistor. However, we will find that different methods can yield significantly different results, which leads to completely different conclusions for the technology under investigation.

This chapter investigates and compares the most commonly used extraction methods. This is done in relation to the above mentioned requirements. The first section of this chapter introduces the extraction methods under investigation. Section 3.2 explains the experimental setup

and the applied criteria. The actual comparison between the methods is made in section 3.3. Finally, section 3.4 discusses issues, which might start to affect mismatch parameter extraction for future technologies. Section 3.5 concludes this chapter.

3.1 Extraction methods

Methods to extract the mismatch in threshold voltage (ΔV_T), current factor ($\Delta\beta/\beta$) or any other transistor parameter (ΔP), can be divided into two groups. The methods in the first group extract transistor parameters for each transistor separately and subtract the results ($\Delta P = P_2 - P_1$). Commonly used methods are:

- Maximum slope method (or steepest slope method)
- Three points method [71]
- Four points method [72]
- Applying a current criterion (only for threshold voltage)

The methods in the second group extract the mismatch ΔP directly by a fit to $\Delta I_D/I_D - V_{GS}$-curves (e.g. see chapter 2). These methods are called *current-mismatch-fitting methods*. The examined methods will now briefly be described.

Maximum slope method. The drain current is measured as a function of the gate bias in the linear regime (low drain bias). The tangent is taken at the place were the steepest slope (g_{mmax}) occurs. The current factor is equal to $\beta = g_{mmax}/V_{DS}$. The gate bias where the tangent and $I_D = 0$ V intercept is equal to $V_T + V_{DS}/2$. The maximum slope method is illustrated in figure 3.1. Note that this method is purely intended for extracting parameters. It does not present a model for the drain current as a function of the bias conditions.

Three points method. The drain current is measured at three gate bias points in strong inversion in the linear regime. The first point is roughly located at maximum transconductance, the second bias point 100-300 mV higher and the third bias point at 'high' gate bias (see figure 3.2). The threshold voltage, current factor and mobility reduction factor (θ) are extracted by solving the following set of equations, which can be done in an analytical way:

$$I_{Di} = \frac{\beta}{1 + \theta(V_{GSi} - V_T - V_{DS}/2)}(V_{GSi} - V_T - V_{DS}/2)V_{DS}, \qquad i = 1, 2, 3 \tag{3.1}$$

A distinction is made with respect to the way the transistors are biased. 1) The three gate bias points have fixed values (e.g.

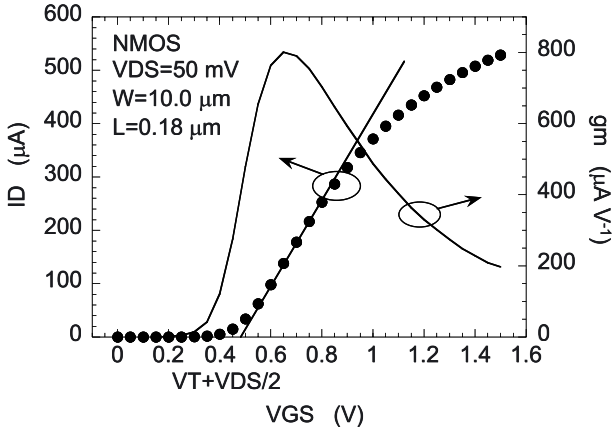
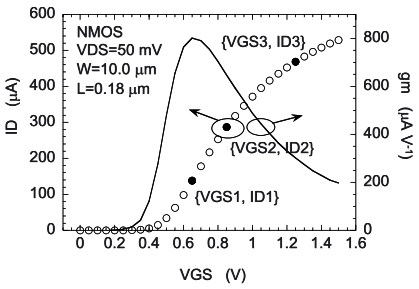
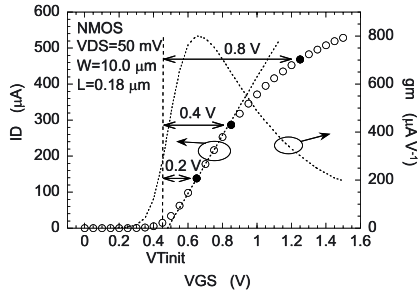


Figure 3.1. Illustration of the maximum slope method



(a) Fixed bias points



(b) Constant overdrive

Figure 3.2. Illustration of the three points method. The solid symbols are measured, open symbols are added for clearness. (a) The three gate-bias conditions have fixed values. (b) The three gate-bias conditions have fixed overdrives with respect to an earlier determined threshold voltage.

$\{V_{GS1}, V_{GS2}, V_{GS3}\} = \{0.7, 0.9, 1.3\}$ V, see figure 3.2a). 2) The gate is biased with a fixed overdrive with respect to an initial threshold voltage (e.g. $\{V_{GS1}, V_{GS2}, V_{GS3}\} = V_{Tinit} + \{0.2, 0.4, 0.8\}$ V, see figure 3.2b). The initial threshold voltage can be determined with any suitable extraction method, including the three points method itself. A couple of iteration cycles can be used in which the latest obtained threshold voltage is taken as the initial threshold voltage for the next cycle. The method that uses fixed overdrive voltages has the advantage that it will yield good results, also when the threshold voltage is, a-priori, not very well known. A disadvantage is the increase in measurement time.

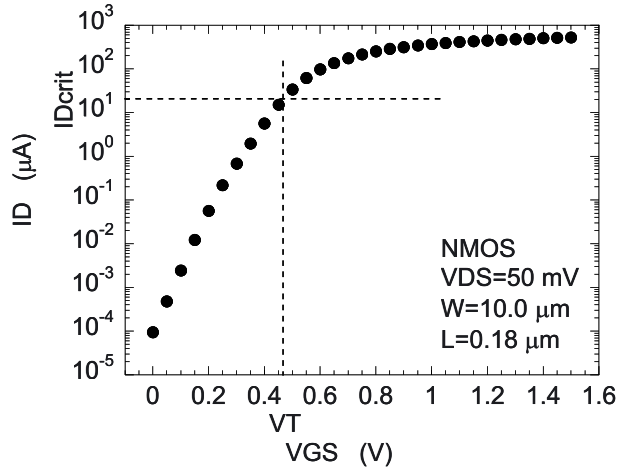


Figure 3.3. Illustration of applying a current criterion to obtain the threshold voltage. An interpolation algorithm is used to extract the correct gate bias.

Four points method. This method is similar to the three points method. However, a more accurate drain current model is used that also takes second order mobility reduction into account:

$$I_{Di} = \frac{\beta(V_{GSi} - V_T - V_{DS}/2)V_{DS}}{1 + \theta_1(V_{GSi} - V_T - V_{DS}/2) + \theta_2(V_{GSi} - V_T - V_{DS}/2)^2} \qquad (3.2)$$

$$i = 1, \cdots, 4.$$

The extra parameter (θ_2) requires one extra measurement point. This set of equations cannot be solved in an analytical way. Therefore, the use of a numerical optimization routine is necessary.

Applying a current criterion. The threshold voltage is defined as the gate bias at a certain current level (I_{Dcrit}). Experimentally obtained current levels at which good values for the threshold voltage are found are $(W/L)*400$ nA for NMOS and $(W/L)*100$ nA for PMOS transistors. Several options are available for finding the correct gate bias. 1) The measurement equipment itself can search for the gate bias, belonging to a particular drain current. Though accurate, this approach can be slow. 2) Another option is to measure the full $I_D - V_{GS}$-curve and use an interpolation algorithm. This is illustrated in figure 3.3. 3) The fastest approach is to connect the gate to the drain and to force I_{Dcrit} into the drain ($I_G \ll I_D$), which gives $V_{GS} = V_{DS} = V_T$. However, this way does not allow for evaluating the drain-bias dependence of threshold-voltage mismatch. In this work the interpolation approach is used, because it

was most easy to implement. As for the maximum slope method, this method is purely intended for extracting parameters. It does not present a model.

Current-mismatch-fitting methods. An explanation of current-mismatch-fitting methods was presented in section 2.3. The mismatch in the drain current ($\Delta I_D/I_D$) is a linear function of the mismatch in the threshold voltage (ΔV_T) and current factor ($\Delta\beta/\beta$). This function is obtained by using a first-order Taylor expansion. The mismatch parameters are extracted by means of a linear least squares fit to experimental $\Delta I_D/I_D - V_{GS}$ curves. Two of these methods will be investigated. The first was published in [31], and will be referred to as fitting method A. The model that is fitted is very similar[1] to:

$$\frac{\Delta I_D}{I_D} = \frac{\Delta\beta}{\beta} - \frac{g_m}{I_D}\Delta V_T \tag{3.3}$$

The fitting range starts at the gate bias where maximum transconductance occurs (in the linear region) and ends at $V_{DS} = V_{DD} = 1.5$ V. No weight is attributed to the measurement points. The second method under investigation is the one described in chapter 2, which will be referred to as fitting method B.

3.2 Experimental setup

In this section the experimental background is provided for comparing the examined extraction methods. Decisions are made concerning: used technology, type of measured devices, geometries of examined device pairs, the amount of measured pairs, test-structure layout, what to measure and data filtering.

Technology. The technology chosen for this experiment is the 0.13 μm technology published in [73], that has a physical oxide thickness of 2.0 nm and a supply voltage of $|V_{DD}| = 1.5$ V. At the time of this work, this experimental technology was stable enough to obtain consistent and relevant results, while for matching studies the technology was advanced. Issues that are not yet important for this technology, but that might appear for future technologies, are described in section 3.4.

Type of devices. In this chapter, only results for NMOS transistors are presented. Similar results were obtained for PMOS transistors.

Device pair geometries. In order to limit measurement time only a selective, but representative, set of transistor pair dimensions was measured, namely:

[1] As opposed to [31], we have taken the liberty of using (2.7) instead of modeling g_m/I_D. See subsection 2.3.2.

W (μm)	L (μm)	
0.25	0.18	narrow and short
10.0	0.18	wide and short
1.0	1.0	wide and long
0.25	7.2	narrow and long
10.0	7.2	large area

Amount of measured pairs. The sample size for this experiment is 84 device pairs per geometry. This results in a relative accuracy of the extracted standard deviations of $\sigma_{\sigma\Delta P}/\sigma_{\Delta P} = 1/\sqrt{2N_{dev}} = 7.7\%$.

Test-structure layout. The most frequently used common source/gate/bulk, separate drains layout was chosen as test structure (see figure 2.2 and figure 2.3a).

What to measure. For each transistor two $I_D - V_{GS}$-curves are measured, one in the linear regime at $V_{DS} = 50$ mV and one in saturation at $V_{DS} = V_{DD} = 1.5$ V. The gate bias ranges from 0.0 to 1.5 volts in steps of 50 mV. All parameters are extracted from the same measured curves. This avoids artifacts in the comparisons due to device or bonding pad degradation. It can occur that an extraction method requires measurements at bias conditions that are not measured, e.g. in the case of the fixed-overdrive three-points method or when applying a current criterion. In these cases an interpolation algorithm is applied on the base curves in order to get the needed 'measurement' data.

Data filtering. Although the investigated technology was stable, at the time of this work it was also still in an experimental phase. Therefore, there might still be some yield issues that can cause extreme parameter shifts for a small amount of the measured devices, that are not related to microscopic fluctuations. To filter out these outliers, a 3σ criterion is applied to the extracted parameters, which is repeated until no more outliers are observed. It might happen that for one extraction method a certain device pair falls just outside the 3σ interval, while for another method it would be just on the inside of it. This could lead to erroneous conclusions when comparing methods. To avoid this problem, a device pair is removed from all data sets when it is considered as an outlier in one of them.

3.3 Comparison of extraction methods

This section compares the extraction methods, that were presented in section 3.1. The methods will be compared with respect to model accuracy (subsection 3.3.1), measurement accuracy (subsection 3.3.2) and

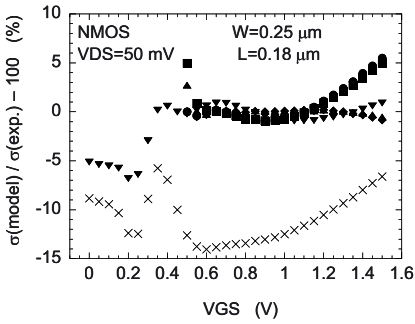
physical meaningfulness of parameters (subsection 3.3.3). The obtained results will be summarized in subsection 3.3.4.

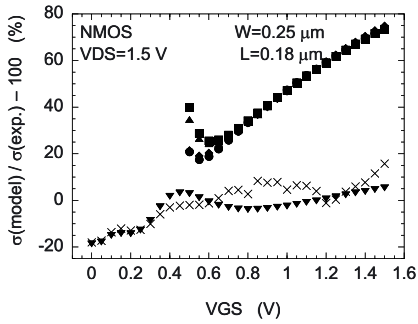
3.3.1 Model accuracy

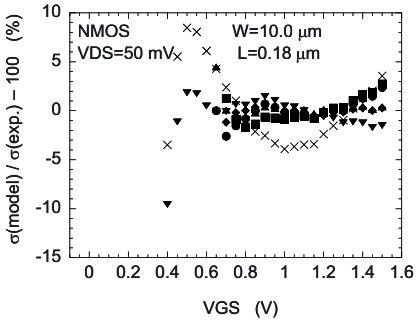
To examine the model accuracy, predicted $\sigma_{\Delta I_D/I_D}$-curves are compared with the experimental curves, that were used to extract the parameters from (see e.g. section 2.3.6). The results of this exercise are presented in figure 3.4. In discussing the results we will distinguish between the direct extraction methods and the current-mismatch-fitting methods.

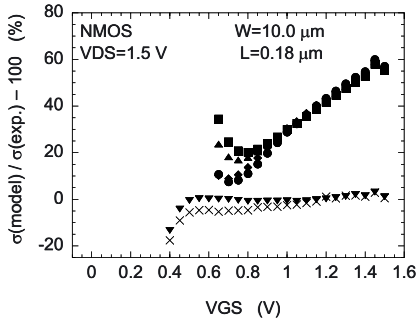
Direct extraction methods. With respect to model accuracy the examined direct extraction methods are limited to the three and four points methods. The maximum slope method and current criterion method are not based on a complete description of the drain current and are therefore disregarded. For the direct methods parameters are extracted in the linear region at low drain bias (V_{DS}=50 mV). It is seen that at this bias condition all methods yield satisfactory accuracy ($|\sigma_{model}/\sigma_{experimental} - 100\%| < 20\%$). At higher gate biases the four points method gives a higher accuracy. This means that second order mobility reduction (the θ_2 term in (3.2)) is present. However, its impact is not big enough to necessitate four points extraction.
Mostly, transistors in analog circuits are not biased in the linear regime but in the saturation regime. The models out of (3.1) and (3.2) can be extrapolated to the saturation region by replacing V_{DS} by the saturation voltage, which is calculated by putting $dI_D/dV_{DS} = 0$. It is seen, however, that this prediction of the mismatch in saturation does not yield accurate results. This is most apparent for short transistor pairs. It can therefore be concluded that effects like velocity saturation and drain induced barrier lowering have to be taken into account. Finally notice that the models out of (3.1) and (3.2) are only valid in strong inversion. Therefore, the weak inversion region was disregarded in the analysis.

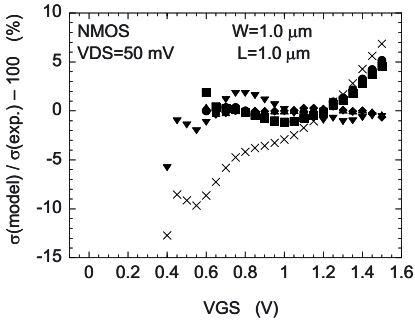
Current-mismatch-fitting methods. In figure 3.4 it is seen that both examined current-mismatch-fitting methods yield good accuracy in as well the linear as the saturation region. Method A gives the highest accuracy. This method extracts a different set of parameters (ΔV_T and $\Delta\beta/\beta$) for the linear region and for the saturation region. It is therefore not continuous between both regions. The four parameters of method B (ΔV_T, $\Delta(1/\beta_0)$, $\Delta(1/\zeta_{sr})$) and $\Delta(1/\zeta_{sat})$ are the same in both regions of operation. This results in continuity over the whole bias range, but at the cost of some accuracy. In weak inversion both models become inaccurate. This is due to the fact that the physical mechanisms, which determine

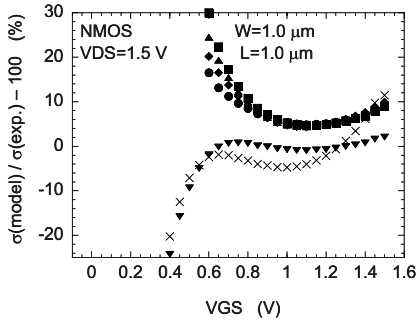

NMOS
VDS=50 mV
W=0.25 µm
L=0.18 µm
σ(model) / σ(exp.) – 100 (%)
VGS (V)

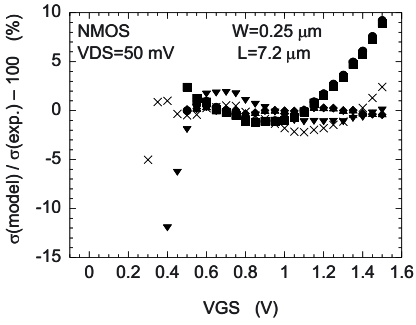

NMOS
VDS=1.5 V
W=0.25 µm
L=0.18 µm
σ(model) / σ(exp.) – 100 (%)
VGS (V)

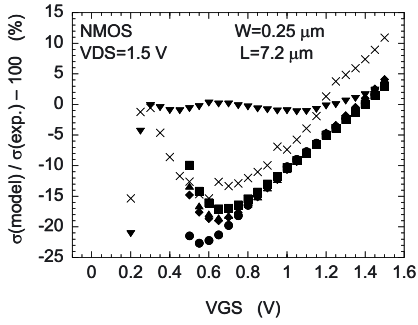

NMOS
VDS=50 mV
W=10.0 µm
L=0.18 µm
σ(model) / σ(exp.) – 100 (%)
VGS (V)


NMOS
VDS=1.5 V
W=10.0 µm
L=0.18 µm
σ(model) / σ(exp.) – 100 (%)
VGS (V)


NMOS
VDS=50 mV
W=1.0 µm
L=1.0 µm
σ(model) / σ(exp.) – 100 (%)
VGS (V)


NMOS
VDS=1.5 V
W=1.0 µm
L=1.0 µm
σ(model) / σ(exp.) – 100 (%)
VGS (V)


NMOS
VDS=50 mV
W=0.25 µm
L=7.2 µm
σ(model) / σ(exp.) – 100 (%)
VGS (V)


NMOS
VDS=1.5 V
W=0.25 µm
L=7.2 µm
σ(model) / σ(exp.) – 100 (%)
VGS (V)

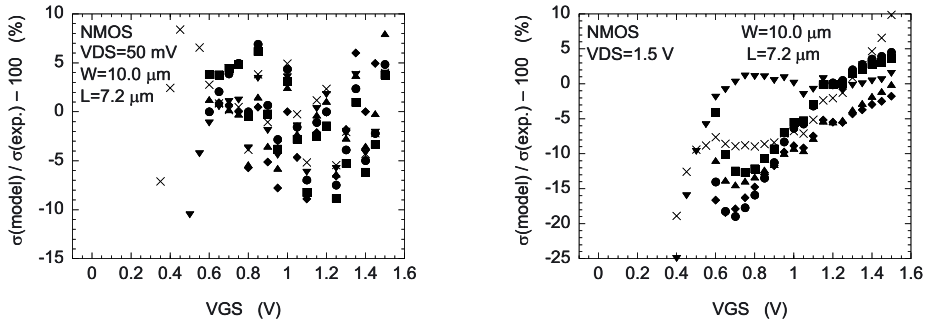


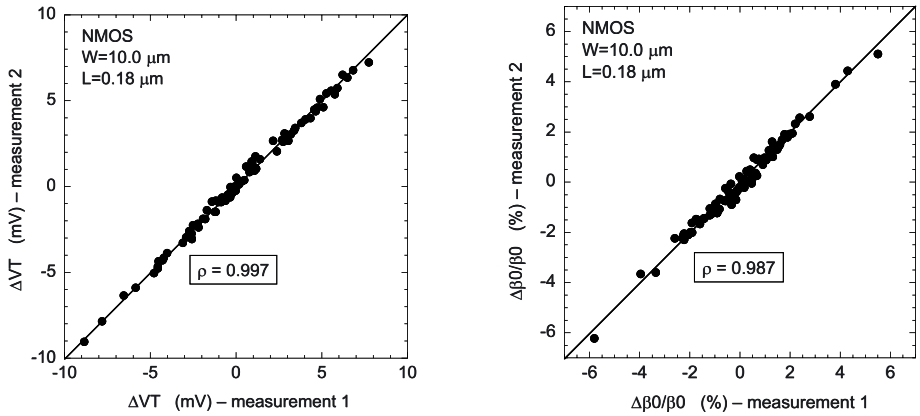
Figure 3.4. Model accuracy for several device-pair dimensions in the linear region (V_{DS} = 50 mV and in the saturation region (V_{DS} = 1.5 V. (●) three points method with fixed bias conditions, (■) three points method with fixed gate overdrive, (⧫) four points method with fixed bias conditions, (▲) four points method with fixed gate overdrive, (▼) current-mismatch-fitting method A, (×) current-mismatch-fitting method B.

threshold voltage mismatch, differ in strong and weak inversion. This will be further looked into in subsection 3.3.3 and in chapter 4.

3.3.2 Measurement accuracy and speed

To determine transistor mismatch, the almost equal drain currents (or other transistor related quantities) of two transistors are subtracted. However, the noise related to the two observations does not cancel out, but is additive. Therefore, determining transistor mismatch requires a much higher measurement accuracy than other transistor measurements. This measurement accuracy is related to required measurement time, since measurement noise can be averaged out by using longer integration times. This subsection mainly deals with measurement accuracy. Measurement speed will be addressed briefly at the end of the subsection.

To examine the measurement accuracy, all measurements were repeated. The second measurement was done at a later time. This means that after the first measurement the wafer was removed from the system. Before the second measurement it had to be reinserted and realigned. Ideally the two measurements should yield exactly the same results. By comparing the extracted parameters of the first and second measurement, like in figure 3.5, the inaccuracy can be determined. As a figure of merit the correlation coefficient between the two measurements ($\rho(\Delta P_1, \Delta P_2)$)



(a) Threshold-voltage mismatch (b) Current-factor mismatch

Figure 3.5. Two examples of measurement repeatability. The method used for extracting the parameters is the three points method with fixed gate overdrive.

is used, which is defined as:

$$\rho(\Delta P_1, \Delta P_2) = \frac{\overline{(\Delta P_1 - \overline{\Delta P_1}) \cdot (\Delta P_2 - \overline{\Delta P_2})}}{\sigma_{\Delta P_1} \cdot \sigma_{\Delta P_2}}. \tag{3.4}$$

Now assume that the measurement result (ΔP) can be split up in the part that we want to measure ($\Delta P_{mismatch}$), which does not change significantly in time, and an unwanted part related to measurement inaccuracies (ΔP_{noise}), i.e. $\Delta P_1 = \Delta P_{mismatch} + \Delta P_{noise1}$. When we further assume that $\rho(\Delta P_{mismatch}, \Delta P_{noise}) = 0$, $\rho(\Delta P_{noise1}, \Delta P_{noise2}) = 0$ and that $\sigma_{\Delta P_{noise}}$ is time invariant , it easily follows that (3.4) calculates the part of $\sigma^2_{\Delta P}$ that is caused by the actual mismatch. The other part is attributed to noise and fluctuations in the resistance between the probe tip and the bonding pad, which will be called contact resistance fluctuations in the remainder of this chapter. Note that one extraction method can be more susceptible to measurement noise than the other.
Tables 3.1 to 3.3 list the correlation coefficients for all examined methods and device dimensions. It is observed that the repeatability is almost 100 % in most cases. The 0.25 μm wide, 0.18 μm long device pairs show the best measurement repeatability, since their intrinsic mismatch is highest. The three and four points methods show poor repeatability on the 10.0 μm wide, 7.2 μm long device pairs. The intrinsic mismatch of these device pairs is low, since they have a large area. This means that they are most susceptible to measurement noise. It will now be shown that this noise is added by the measurement equipment. Figure 3.6 displays, on the right axis, the measurement repeatability (ρ_{repeat})

Table 3.1. Correlation coefficients between two measurements of ΔV_T, evaluating measurement repeatability. The device width/length ratios at the top of the columns is given in μm/μm. The abbreviation f.b. stands for fixed bias conditions and f.o. stands for fixed gate bias overdrive.

model	V_{DS} (V)	0.25/0.18	10.0/0.18	1.0/1.0	0.25/7.2	10.0/7.2
maximum slope	0.05	0.99916	0.99733	0.99765	0.99825	0.99465
3 points – f.b.	0.05	0.99903	0.99591	0.99648	0.99521	0.55807
3 points – f.o.	0.05	0.99928	0.99688	0.99715	0.99719	0.72763
4 points – f.b.	0.05	0.99678	0.98497	0.98824	0.98220	0.61809
4 points – f.o.	0.05	0.99823	0.99140	0.99061	0.98264	0.58814
current criterion	0.05	0.99989	0.99987	0.99979	0.99958	0.99985
current criterion	1.5	0.99981	0.99977	0.99972	0.99963	0.99942
fitting method A	0.05	0.99989	0.94600	0.99984	0.99976	0.99453
fitting method A	1.5	0.99985	0.99984	0.99969	0.99943	0.99960
fitting method B	both	0.99977	0.99888	0.99964	0.99925	0.99884

Table 3.2. The same as table 3.1, but now for $\Delta\beta/\beta$

model	V_{DS} (V)	0.25/0.18	10.0/0.18	1.0/1.0	0.25/7.2	10.0/7.2
maximum slope	0.05	0.99973	0.98936	0.99870	0.99785	0.99348
3 points – f.b.	0.05	0.99961	0.99100	0.99751	0.99440	0.49438
3 points – f.o.	0.05	0.99951	0.98707	0.99702	0.99587	0.54931
4 points – f.b.	0.05	0.99721	0.94537	0.97527	0.96600	0.36262
4 points – f.o.	0.05	0.99748	0.94604	0.97758	0.96668	0.37339
fitting method A	0.05	0.99999	0.97170	0.99995	0.99996	0.99578
fitting method A	1.5	0.99998	0.99973	0.99995	0.99982	0.99962
fitting method B	both	0.99979	0.99743	0.99965	0.99804	0.99714

of threshold voltage mismatch as a function of the placement of the three bias points[2]. Also shown, on the left axis, is the drain current. It is seen that the repeatability drops significantly when, at the second bias point (V_{GS2}), the measurement system switches to a higher current-

[2]Figure 3.6 presents results for the three points method with fixed bias points. Similar figures could have been made for the other three and four points methods.

Table 3.3. The same as table 3.1, but now, depending on the method, for $\Delta\theta$, $\Delta\theta_1$, $\Delta\theta_2$, $\Delta(1/\zeta_{sr})$ or $\Delta(1/\zeta_{sat})$

model	parameter	0.25/0.18	10.0/0.18	1.0/1.0	0.25/7.2	10.0/7.2
3 points – f.b.	$\Delta\theta$	0.99849	0.94844	0.99551	0.99437	0.22099
3 points – f.o.	$\Delta\theta$	0.99774	0.93861	0.99480	0.99503	0.31691
4 points – f.b.	$\Delta\theta_1$	0.98863	0.87339	0.96970	0.96490	0.21776
4 points – f.b.	$\Delta\theta_2$	0.98464	0.82076	0.94668	0.93615	0.18784
4 points – f.o.	$\Delta\theta_1$	0.98908	0.84834	0.96785	0.96549	0.16560
4 points – f.o.	$\Delta\theta_2$	0.98502	0.76111	0.94308	0.93834	0.10136
fitting method B	$\Delta(1/\zeta_{sr})$	0.99874	0.96694	0.99964	0.99665	0.99184
fitting method B	$\Delta(1/\zeta_{sat})$	0.99947	0.96949	0.99379	0.98782	0.99677

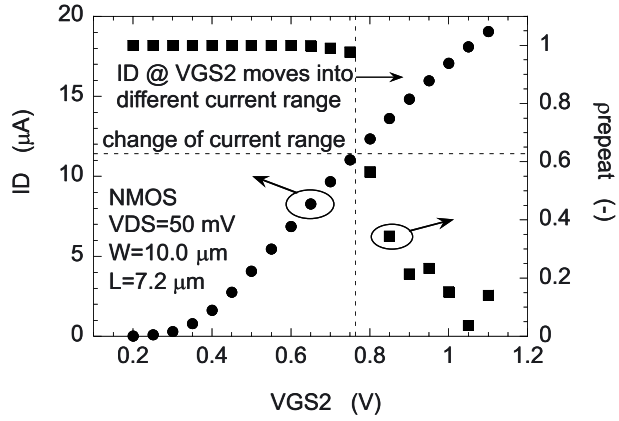


Figure 3.6. Measurement repeatability of ΔV_T (right axis) as a function of the placement of the second bias points for the three points method with fixed bias conditions. Also shown is the drain current (left axis) and where the measurement system switches from current range (dashed lines). $V_{GS1} = V_{GS2} - 0.2$ V, $V_{GS3} = V_{GS2} + 0.4$ V

measurement range. At the low end of a measurement range, the system noise is highest, which, in this case, demonstrates the impact of the system. To obtain more accurate measurements, longer integration times would be necessary. Note that the other methods do not suffer from the added noise. For the fitting methods, the noise is averaged out over the large number of measured bias points. The maximum slope method does not suffer in this particular case, because the peak transconductance is still located in the high end of the lower current range. If this would not have been the case, this method would also have shown poor

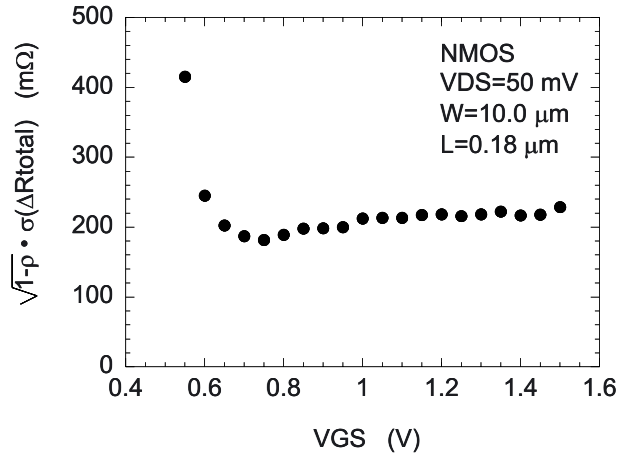


Figure 3.7. Non-repeatable part of the total measured resistance fluctuations of the 10.0 μm wide, 0.18 μm long device pairs at $V_{DS} = 50$ mV

measurement repeatability. Using a current criterion gives a very good measurement repeatability, since the intrinsic mismatch of a device pair is much higher around threshold than it is in strong inversion (see e.g. section 2.3 or figure 2.10).
Returning to table 3.3, it can be seen that the 10.0 μm wide, 0.18 μm long device pairs have a slightly worse repeatability for $\Delta\theta$ (or other related parameters) than the other pair dimensions. This is believed to be due to contact resistance fluctuations. These device pairs are most susceptible to these fluctuations since they are wide and short and therefore have low resistance by themselves. Figure 3.7 plots the non-repeatable part of the total measured resistance ($R_{total} = V_{DS}/I_D$) fluctuations in the linear region, which, as expected, are constant as a function of the gate bias. From this, the contact resistance fluctuation is calculated to be $\sigma_{R_{contact}} \approx 150$ mΩ per bonding pad. This number can easily become larger when bonding pads are degraded by earlier measurements, the probes themselves are worn-out or when the pressure of the probes on the bonding pads is too low. In the next subsection it will be seen that contact resistance can affect current factor mismatch when series resistance is not included in the analysis. Contact resistance fluctuations are only an issue for measurements in the linear region. In the saturation region the dependence of the drain current on the drain bias is weak and the influence of series resistance at the drain diminishes. Series resistance at the source still impacts the drain current, but it is common to both transistors in the pair and therefore does not affect the mismatch. Another source of inaccuracy, which was not observed in this particular

experiment, could be the measurement resolution of the system. Currents are presented at a 5.5 digits resolution. Current factor mismatch is approximately equal to $(1\ \%\mu\mathrm{m})/\sqrt{WL}$. This means that measurement resolution could start to play a role for transistors with an area larger than 100 $\mu\mathrm{m}^2$.

A priori, it is difficult to put exact numbers to when matching measurements start to fail. The specified worst-case system accuracy for current measurements (see section 2.1.1) is about 0.5 %. This suggests that transistors with an area of 1 $\mu\mathrm{m}^2$ or larger cannot accurately be measured. This statement is clearly contradicted by tables 3.1 to 3.3. Matching measurements are relative measurements by nature. Their accuracy is not determined by the long-term worst-case system accuracy, but by the short-term system repeatability, which is much better. Usually, this parameter is not specified by the equipment vendor. To further complicate issues, measurement repeatability can depend strongly on the device under test, as was observed in figure 3.6. In this example, for somewhat narrower devices the first two bias points would have fallen inside the more accurate lower current range, with almost 100 % repeatability for ΔV_T measurements. It can be concluded that accurate matching measurements are possible far beyond the specified system accuracy. However, one has to always remain careful, especially for transistors with large area ($\gtrsim 10\ \mu\mathrm{m}^2$, due to system noise) or for short, wide transistor pairs ($\sigma_{\Delta R} \lesssim 2\ \Omega$, due to problems with contacting).

We will end this subsection with a discussion on measurement speed. The time a measurement takes is roughly proportional to the number of bias conditions needed for the extraction.[3] Listing the different methods from the fastest to the slowest gives: 1) three points method with fixed bias conditions (∼0.9 s[4]), 2) four points method with fixed bias conditions (∼1.2 s), 3) three points method with fixed gate bias overdrive (∼1.8 s), 4) four points method with fixed gate bias overdrive (∼2.4 s), 5) applying a current criterion (∼2 s), 6) maximum slope method (∼3 s), 7) current-mismatch-fitting methods A (∼7.5 s), and 8) current-mismatch-fitting method B (∼15 s). From the perspective of measurement time, using a three or four points method is preferred. However, as was concluded earlier, for large area transistors one has to be careful about measurement noise.

[3] in section 2.1.1 it was found that the measurement of one bias point requires approximately 300 ms

[4] This is the time it approximately takes to measure one transistor pair

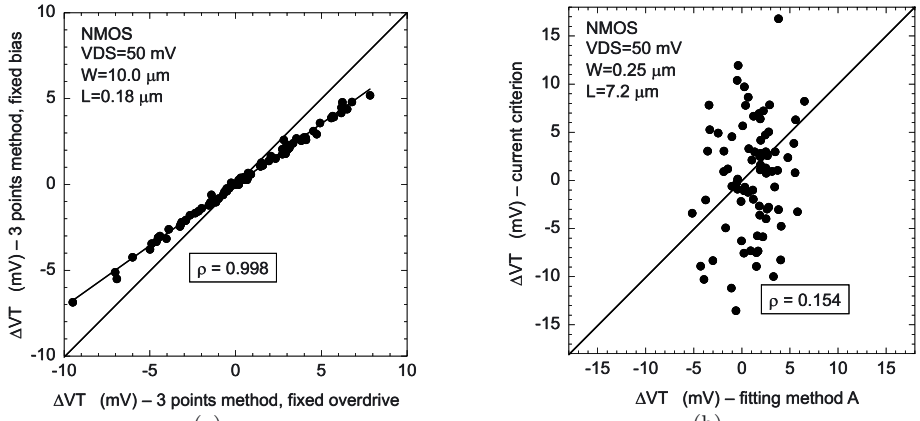


Figure 3.8. Two examples, where threshold-voltage-mismatch-extraction methods are compared. In (a) the three points method with fixed gate overdrive is compared to the same method using fixed bias conditions. In (b) current-mismatch-fitting method A is compared to applying a current criterion.

3.3.3 Physical meaningfulness of parameters

To investigate whether parameters, that are extracted by two different methods, have the same physical meaning, a device pair by device pair comparison is made. Two examples are shown in figure 3.8, which plot ΔV_T extracted with one method against ΔV_T extracted with a different method. Two kinds of differences can be observed. In the first case (figure 3.8a) the correlation is almost 1, but a difference in slope occurs, i.e. $\Delta V_{T,method1} = f \cdot \Delta V_{T,method2}$ is noticed. This means that one (or both) of the methods systematically under-/overestimates the mismatch. However, because of the high correlation, it can be concluded that the mismatch causing effect is the same for both situations. This does not have to be the case, as is illustrated in figure 3.8b. Note that both methods have almost 100 % measurement repeatability and that mismatch parameters are extracted from the same measurement curves. The poor correlation is therefore truly caused by a difference in physical content of the parameters. Using (3.4), it can easily be shown that the correlation between two parameters ($\rho(\Delta P_1, \Delta P_2)$) cannot be purely explained by the measurement inaccuracy of both parameters ($\rho_{repeat}(\Delta P_1)$ and $\rho_{repeat}(\Delta P_1)$) when:

$$\rho(\Delta P_1, \Delta P_2) < \sqrt{\rho_{repeat}(\Delta P_1) \cdot \rho_{repeat}(\Delta P_2)}. \tag{3.5}$$

In summary, by comparing $\sigma_{\Delta P}$ values, it can be determined whether the absolute value of the extracted threshold voltage mismatch is correct. By correlating different methods with each other, the physical content of the parameters can be examined. Another way to examine whether an extracted parameter has the expected physical content is to introduce a known mismatch in the devices under test and to look at how well this is reproduced by the extraction method. For current factor mismatch this could be achieved by using a dedicated test structure, which had a designed systematic mismatch in the gate lengths of the two transistors in the pair [74].
Tables 3.4 to 3.6 show the results of the above mentioned tests. The standard deviations of threshold-voltage and current-factor mismatch for the examined extraction methods are listed in the top parts of tables 3.4 and 3.5 respectively. The bottom parts of these tables list the correlation of the examined methods with the three points method with fixed gate bias overdrive, which has been chosen as the method of reference. Table 3.6 lists the medians of extracted current-factor mismatch for small introduced mismatch in the gate length (top part of the table) and a larger introduced mismatch (bottom part of the table) for several average gate lengths. As an estimate, the median was preferred over the mean, since the first is less sensitive to outliers.
We will now discuss the results presented in these tables method by method.

Maximum slope method. Looking at tables 3.4 and 3.5 it is observed that the maximum slope method gives results that are close to the results of the three- and four-points methods with fixed gate bias overdrive. Significant differences are only observed for the 10.0 μm wide, 7.2 μm long device pairs, that are caused by poor measurement repeatability (see tables 3.1 and 3.2).
Looking at the bottom part of table 3.6 it is seen that the maximum slope method slightly underestimates the current factor mismatch. This underestimation becomes more prominent for decreasing transistor lengths. This suggests that it is related to the influence of series resistance or other short channel phenomena. The decrease is in contradiction with what one would expect. The introduced mismatch is equal to $\Delta\beta/\beta = -\Delta L/L_{mask}$. However, the physical mismatch is not proportional to the inverse of the mask length, but to the inverse of the smaller effective channel length. Understanding the observed behavior is difficult. The position and height of the maximum transconductance peak originate from a mixture of physical effects ranging from mobility reduction, series resistances effects and gate depletion [21] to the usually poorly described transitions from saturation to the linear region and

Table 3.4. Standard deviations of ΔV_T in mV and the correlation of ΔV_T with the ΔV_T's extracted with the three points method with fixed gate overdrive. The device width/length ratios at the top of the columns are given in μm/μm. The abbreviation f.b. stands for fixed bias conditions and f.o. stands for fixed gate overdrive.

model	V_{DS} (V)	0.25/0.18	10.0/0.18	1.0/1.0	0.25/7.2	10.0/7.2
			$\sigma_{\Delta V_T}$ (mV)			
maximum slope	0.05	17.896	3.5518	4.9085	2.9830	0.59237
3 points – f.b.	0.05	13.167	2.5999	3.7867	2.1091	0.66291
3 points – f.o.	0.05	18.130	3.6504	4.9146	2.7862	0.78975
4 points – f.b.	0.05	13.707	2.8391	4.4826	3.1177	0.87543
4 points – f.o.	0.05	18.358	3.6700	5.1031	3.2245	0.85381
current criterion	0.05	21.116	3.6652	7.0361	5.9258	1.0062
current criterion	1.5	21.190	4.4757	8.8511	5.3200	1.1544
fitting method A	0.05	20.994	3.6004	6.1162	2.5110	0.50281
fitting method A	1.5	17.572	4.0010	4.7539	3.7193	0.74600
fitting method B	both	18.752	4.3143	5.6053	4.5970	0.99666
correlation with three points method with fixed gate bias overdrive (-)						
maximum slope	0.05	0.97986	0.98987	0.98211	0.97021	0.81125
3 points – f.b.	0.05	0.99550	0.99771	0.99643	0.99737	0.97027
4 points – f.b.	0.05	0.91693	0.96990	0.94879	0.95381	0.85954
4 points – f.o.	0.05	0.95891	0.98302	0.96107	0.95738	0.54038
current criterion	0.05	0.6547	0.91925	0.63421	0.51844	0.70807
current criterion	1.5	0.57393	0.82957	0.49337	0.38342	0.44883
fitting method A	0.05	0.83422	0.97520	0.71726	0.57656	0.56195
fitting method A	1.5	0.80575	0.91610	0.70299	0.55112	0.61002
fitting method B	both	0.88373	0.92030	0.70736	0.63438	0.61672

from weak to strong inversion.
Now consider the top part of table 3.6, that shows the results for the case of small introduced mismatch. For two of the examined pair dimensions it is observed that the maximum slope method gives serious overestimation of the systematic mismatch, which is caused by a bad contact of a probe-needle at one of the drains. The overestimation does not occur for the extraction methods that take series resistance into account, because it is filtered out by the mobility reduction parameters.
In conclusion, the maximum slope method extracts consistent values for the standard deviation of threshold-voltage mismatch and current-factor mismatch. However, it is sensitive to series resistance effects. Another

Table 3.5. The same as table 3.4, but now for $\Delta\beta/\beta$

model	V_{DS} (V)	0.25/0.18	10.0/0.18	1.0/1.0	0.25/7.2	10.0/7.2
		$\sigma_{\Delta\beta/\beta}$ (mV)				
maximum slope	0.05	15.069	1.2235	2.6847	0.99026	0.23389
3 points – f.b.	0.05	14.444	1.4531	2.4897	0.94718	0.38996
3 points – f.o.	0.05	14.645	1.3744	2.5875	1.1001	0.41496
4 points – f.b.	0.05	13.879	1.5731	2.2272	1.3253	0.38996
4 points – f.o.	0.05	13.914	1.5022	2.3644	1.3635	0.51299
fitting method A	0.05	16.923	1.3085	3.1844	1.0431	0.23982
fitting method A	1.5	10.215	0.82103	2.8676	0.90693	0.21956
fitting method B	both	14.491	1.1971	3.0085	1.7888	0.39968
correlation with three points method with fixed gate bias overdrive (-)						
maximum slope	0.05	0.96473	0.91485	0.94897	0.90832	0.70044
3 points – f.b.	0.05	0.97726	0.90439	0.97266	0.96199	0.95784
4 points – f.b.	0.05	0.89365	0.80661	0.90044	0.88517	0.77524
4 points – f.o.	0.05	0.93988	0.90167	0.90937	0.88678	0.23759
fitting method A	0.05	0.88765	0.76998	0.81090	0.50186	0.58842
fitting method A	1.5	0.80545	0.57800	0.65542	0.45919	0.47079
fitting method B	both	0.91647	0.90506	0.84393	0.69302	0.60929

drawback of the method is that the physical effects, forming the maximum transconductance peak, are not easily modeled.

Three and four points methods. We will start by comparing the three and four points method with fixed gate-bias overdrive. It is observed that these methods yield approximately the same results (see tables 3.4 and 3.5). Although the differences are somewhat larger than those expected from the repeatability study in subsection 3.3.2, they are small enough to disregard. When comparing the methods with fixed gate-bias conditions (f.b.) with the methods with fixed gate-bias overdrive (f.o.) it is observed that the extracted standard deviations differ, despite a high correlation. For example in the case of threshold voltage mismatch, the extracted standard deviations are significantly lower in the case of fixed bias conditions. This stems from the fact that the models represented by (3.1) and (3.2) are not 100 % accurate. Because of this inaccuracy, the extracted parameters are a function of the placement of the bias points, as is illustrated in figure 3.9a for the threshold voltage. This behavior gives rise to a feedback mechanism that causes underestimation when fixed bias conditions are applied. Consider a transistor with a

Table 3.6. Medians of $\Delta\beta/\beta$ in % extracted with the examined methods for device pairs with small intentional mismatch in mask length (-5.71 %) and for device pairs with larger intentional mismatch (-28.57 %). The device width/(average length) ratios at the top of the columns are given in μm/μm. The abbreviation f.b. stands for fixed bias conditions and f.o. stands for fixed gate bias overdrive.

method	V_{DS} (V)	1.5/0.175	4.5/0.525	13.5/1.575	40.5/4.725
		$\Delta\beta/\beta = 5.71$ %			
maximum slope	0.05	14.12	5.20	18.40	5.63
3 points – f.b.	0.05	4.74	5.31	4.10	5.42
3 points – f.o.	0.05	7.72	5.70	5.62	5.61
4 points – f.b.	0.05	4.43	5.32	4.24	5.59
4 points – f.o.	0.05	7.56	5.57	5.52	5.39
fitting method A	0.05	19.55	4.86	31.02	5.61
fitting method A	1.5	3.37	4.24	5.40	5.68
fitting method B	both	5.36	6.37	4.56	5.67
		$\Delta\beta/\beta = 28.57$ %			
maximum slope	0.05	24.17	25.08	26.07	27.78
3 points – f.b.	0.05	29.23	25.42	24.55	26.81
3 points – f.o.	0.05	31.89	27.33	26.33	27.87
4 points – f.b.	0.05	30.42	25.76	24.76	28.07
4 points – f.o.	0.05	33.03	27.60	26.17	26.68
fitting method A	0.05	16.93	23.49	25.81	27.33
fitting method A	1.5	14.61	20.30	25.35	27.16
fitting method B	both	34.18	32.11	28.24	28.63

threshold voltage ΔV_T higher than the average. For not too large ΔV_T, the mismatch is underestimated by a factor $1/(1 - dV_T/dV_{GS1})$. More generally it can be stated that:

$$\Delta P_{f.o.} \cong \Delta P_{f.b.} + \frac{dP}{dV_{GS1}} \Delta V_{T,f.o.}, \tag{3.6}$$

which for the threshold voltage is illustrated in figure 3.9b. In this figure it is also observed that $\sigma_{\Delta V_T}$ depends less strongly on the placement of the measurement points when it is extracted with fixed gate overdrive. Notice that when measurement time is a serious constraint, one could use fixed bias conditions in combination with (3.6), although a few extra measurements would be needed to accurately determine dP/dV_{GS1}.
Now consider the results presented in table 3.6. As expected, it is observed that the intentional systematic mismatch is best reproduced using

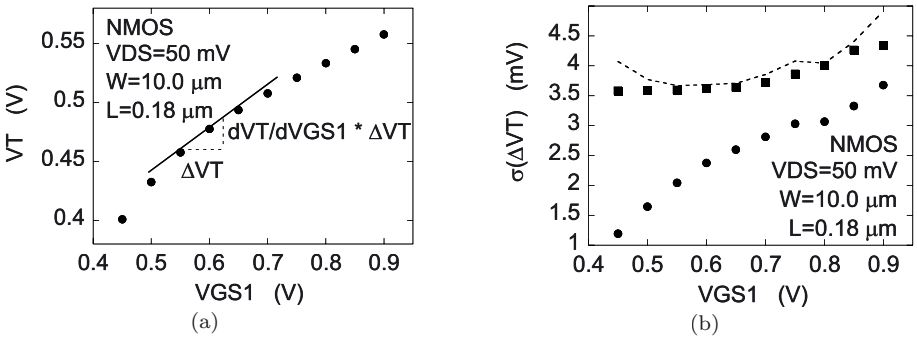
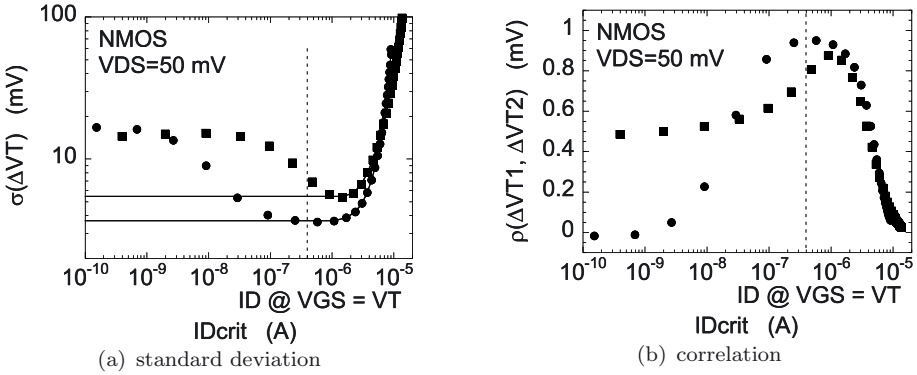


Figure 3.9. Extracted threshold voltage (a) and threshold-voltage mismatch (b) with the three points method as a function of the placement of the bias points. (•) fixed gate bias, (■) fixed gate bias overdrive. The from (3.6) predicted threshold-voltage mismatch is represented by the dashed line. $V_{GS2} = V_{GS1} + 0.2$ V, $V_{GS3} = V_{GS1} + 0.6$ V.

the methods with fixed gate bias overdrive. Looking at the shortest device pairs with small intentional mismatch, one can see that the methods with fixed gate overdrive extract the expected larger mismatch, while the methods with fixed bias conditions yield smaller results. Note that no impact of the bad contact is observed.
In conclusion, when using the three or four points method, the transistors should be biased using fixed gate overdrive. No clear difference between the three and four points method was observed. However, note that the four points method is more difficult to implement, since it requires a numerical optimization algorithm.

Applying a current criterion. In table 3.4 it is observed that using a current criterion leads to larger values for the extracted standard deviation of threshold voltage mismatch than when another method is applied. Also the correlation with the three points method is significantly smaller than 1. To understand this behavior, figure 3.10a plots the extracted standard deviation as a function of the applied current level. Figure 3.10b plots the correlation with the three points method. At large values of the current level the mismatch is dominated by current-factor mismatch (see section 2.3). With decreasing current level the standard deviation is expected to decrease and to level off at $\sigma_{\Delta V_T}$ as is depicted by the solid line. The correlation is expected to increase and to level off at 1. However, experimentally it is observed that when the current level is decreased into the weak inversion region, the standard deviation starts to increase again and the correlation drops. This shows that the physical mechanisms that cause threshold-voltage mismatch are not the same



(a) standard deviation (b) correlation

Figure 3.10. a) Standard deviation of the threshold voltage mismatch extracted with a current criterion as a function of the applied current level. b) Correlation of the threshold voltage mismatch extracted with the three points method with fixed gate overdrive and the mismatch extracted by a current criterion as a function of the applied current level. The solid lines represent the case of equal threshold voltage mismatch in the weak and strong inversion region. (•) $W = 10.0$ μm, $L = 0.18$ μm. (■) $W = 1.0$ μm, $L = 1.0$ μm.

in the weak and strong inversion regions.[5] Applying a current criterion extracts the threshold voltage in the intermediate moderate-inversion region, which leads to results that are difficult to interpret. A reasonable estimate for $\sigma_{\Delta V_T}$ in strong inversion can be obtained by equating it to the minimum of the $\sigma_{\Delta V_T} - I_{Dcrit}$ curve. However, this requires higher current levels, that are, a priori, not known. This leads to an increase in measurement time.

Current-mismatch-fitting method A. Comparing fitting method A with the other methods, it is observed that the correlation is poor (see tables 3.4 and 3.5). This is due to the fact that physical effects such as mobility degradation, series resistance and velocity saturation (in the case of $V_{DS} = 1.5$ V) are not included in the model, while the fit includes bias conditions for which these effects play a role. This conclusion can also be drawn when looking at the systematic current-factor mismatch caused by the intentional mismatch in device length (see table 3.6). The introduced mismatch is seen to be underestimated, which is most noticeable for the short transistor pairs. Note that the extraction at $V_{DS} = 50$ mV suffered from the bad contact of one of the needles. This was already observed for the maximum slope method and can be solved by filtering out series resistance by including a mobility reduction parameter. In

[5]Chapter 4, section 4.2 will look into this difference in more detail.

Table 3.7. Qualitative comparison of the extraction methods with respect to model accuracy, measurement accuracy, sensitivity to contacting errors, measurement speed and physical meaningfulness of parameters. The abbreviation f.b. stands for fixed bias conditions and f.o. stands for fixed gate overdrive.

model	model accuracy	measurement accuracy	sensitivity to contact	measurement speed	physical content
maximum slope	N.A.	-	-	0	0
3 points – f.b.	-	-	+	+	0
3 points – f.o.	-	-	+	0	+
4 points – f.b.	-	-	+	+	0
4 points – f.o.	-	-	+	0	+
current criterion	N.A.	+	+	0	-
fitting method A	+	+	-	-	-
fitting method B	+	+	+	-	0

saturation (V_{DS} = 1.5 V) the impact of bad contacting disappears as was explained in subsection 3.3.2.

Current-mismatch-fitting method B. Finally, looking at tables 3.4 and 3.5, it is observed that fitting method B extracts larger standard deviations of threshold-voltage mismatch and current-factor mismatch than the three or four points methods. The correlation between the methods ranges from 60 % to 92 %. Fitting method B fits a simple model over a large bias range. Modeling errors are averaged out over all parameters. Looking at table 3.6 it is observed that the method gives a reasonable estimate of the intentionally introduced mismatch in gate length.

3.3.4 Summary

To end this section, table 3.7 qualitatively compares the examined extraction methods with respect to model accuracy, measurement accuracy, sensitivity to contacting errors, measurement speed and physical meaningfulness of parameters.

3.4 Future issues

This section will briefly discuss three issues that might start to affect matching measurements for technologies beyond the 0.13 μm node. The first is related to the gate leakage current, the second to problems with contact resistance and the third to the bad description of the mismatch in between weak and strong inversion.

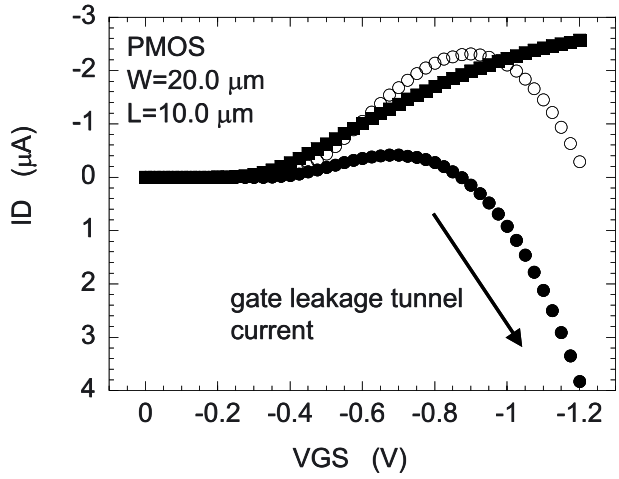


Figure 3.11. Drain current as a function of the gate bias for transistors with thin gate oxide. (•) equivalent $t_{ox} = 1.35$ nm, $V_{DS} = 40$ mV. (○) equivalent $t_{ox} = 1.35$ nm, $V_{DS} = 100$ mV. (■) equivalent $t_{ox} = 1.55$ nm, $V_{DS} = 40$ mV.

Gate leakage current. As the gate oxide thickness scales down, the leakage current due to tunnelling increases. This current is proportional to the transistor length, while the drain current without the leakage contribution is inversely proportional to this length. Also, the leakage current increases exponentially with gate bias, while at low V_{DS}, the drain current increases linearly. Figure 3.11 shows the total measured drain current for 20.0 μm wide, 10.0 μm long transistors for two values of the effective oxide thickness and two values of the drain bias. For the thin gate-oxide transistor it is observed that the leakage component to the drain current becomes dominant at high gate bias. None of the methods that were described in section 3.1 takes this leakage current into account. The methods are only valid for the limited bias range in which the tunnelling current is not significant. To allow parameter extraction outside this range, new models need to be developed, that take the matching properties of the gate tunnelling current into account. The introduction of high-k dielectrics would overcome this problem.

Contact resistance. Effects of contact resistance, though small, were already observed in subsections 3.3.2 and 3.3.3. However, as gate lengths scale down, the resistance of the transistor pairs under test decreases and the measurement problems related to contact resistance could become disastrous. A possible solution would be to decrease the width of the measured transistors, but this would mean that transistors used in RF circuits, wide by nature, cannot be examined directly. A way to circum-

vent the problem would be to use a force and sense technique, where the sensing is done with a different needle and therefore beyond the contact. This would require two connections and bonding pads attached to each drain, one for forcing the drain voltage and the other one for sensing it. This, however, increases the size of the test structure significantly. Note that it is also possible to put two probe-needles on a somewhat larger designed bonding pad. A more elegant solution would be to change the layout of the test structure to a common source, common drain, separate gates configuration, instead of using the standard common source, common gate, separate drains layout. The contact resistance is now common to both transistors and has therefore no significant impact on the mismatch. The contact resistance at the gate is negligible to the large resistance of the gate itself. This adapted layout would also allow investigation of the mismatch in gate current, which was mentioned earlier.

Moderate inversion. As supply voltages scale down, the range in which designers can bias transistors, becomes smaller. This pushes the operating conditions of transistors more and more towards moderate inversion. As was pointed out in subsection 3.3.3 the moderate inversion region is difficult to describe and from matching point of view it is unknown territory. Although the formulation presented in section 2 pushed the model validity to lower values of the gate bias, figure 3.10 clearly shows that more modeling efforts are required to better understand and describe this region.

3.5 Conclusions

In this chapter the most common techniques to extract the mismatch of a pair of MOSFETs have been compared, namely: the maximum slope method, three and four points methods, applying a current criterion and current-mismatch-fitting methods. The comparison was made with respect to model accuracy, measurement accuracy and speed, and physical meaningfulness of extracted parameters.

Regarding model accuracy, it was found that in the saturation region current-mismatch-fitting methods yield the highest accuracy. The examined direct extraction methods only use the linear regime for parameter extraction. In this region the models provide accurate results. However, when extrapolated to saturation the direct methods do not yield accurate results which shows that effects like velocity saturation and drain induced barrier lowering cannot be ignored. Of the examined current-mismatch-fitting methods, the method published in [31] gives the best accuracy. The method developed in chapter 2 has somewhat lower accuracy, but has the advantage that it is continuous from the linear regime

to saturation.
The measurement accuracy related to the extraction methods was examined by means of a repeatability study. Two sources of error have been recognized, errors caused by measurement system noise and errors caused by fluctuations in the resistance between the probe tips and the bonding pads. The three and four points methods, and assumably the maximum slope method, were found to be most sensitive to measurement system noise. Noise was found to only affect transistor pairs with large area since their intrinsic mismatch is small. No clear limit could be determined since the short term repeatability of measurement systems is not specified, while this quantity was also found to depend strongly on the measured current level. When applying a current criterion, no significant impact of measurement noise is observed, because the mismatch in the drain current is large at the low current level at which the extraction takes place. In the case of current-mismatch-fitting methods the noise is averaged out over the large number of measured bias conditions. Although, with respect to measurement noise, one has to be careful when using a three or four points method, they are preferred methods in industrial environments, because of the limited measurement time required.
Contact resistance fluctuations were observed, but they were not large enough to be a limiting factor when extracting standard deviations. An impact of contact resistance was seen on the average current factor mismatch for the maximum slope method and for fitting method A in the linear regime. These methods do not take series resistance into account. In case of the other methods series resistance effects are filtered out by the mobility reduction parameters. Contact resistance does not impact measurement accuracy in the saturation region.
The physical meaningfulness of the extracted mismatch in threshold voltage and current factor has been investigated by comparing the extracted standard deviations and by correlating the results obtained by the different extraction methods. Current-factor mismatch was further investigated by using a dedicated test structure, which has intentional mismatch in the gate length. It was found that the physically most meaningful parameters were obtained by using the three points method or four points method. However, it was seen that the transistors have to be biased using a fixed gate overdrive. Using fixed bias conditions leads to a wrong estimate, caused by the inaccuracy of the models on which the methods are based. Using the maximum slope method also provided good results, but it was found to be sensitive to contact resistance. When investigating the application of a current criterion, it was found that threshold voltage mismatch in weak inversion is caused by

different physical mechanisms than threshold voltage mismatch in strong inversion. Conventional current levels are located in between these two regions, which makes the result difficult to interpret. One either has to use a lower current level to investigate the weak inversion region, or a higher current level for the strong inversion region. The latter can only be applied if current factor mismatch is still negligible at this higher current level. Current-mismatch-fitting methods were seen to provide physically less meaningful parameters. Model inaccuracies related to the simple models on which these methods are based are averaged out over all the extracted parameters.

Finally some issues are recognized that might start to affect parameter extraction for up-coming technologies. Due to the down-scaling of gate thickness, gate leakage starts to have a serious impact on the drain current of long transistors. This means that matching measurements will either be limited to devices with not too long gate lengths or that this leakage current needs to be taken into account. Because of the down-scaling of the gate length, problems with contacting are expected to increase, which will limit matching measurements to not too wide transistors. To circumvent this problem it has been proposed to change the test-structure layout for matching measurements to a common source, common drain, separate gates configuration. Finally, as supply voltages scale down, analogue operation of the MOSFET is pushed towards the moderate inversion region. A modeling effort is required to better understand this region of interest.

Chapter 4

PHYSICAL ORIGINS OF MOSFET MISMATCH

In the previous two chapters it was examined how MOSFET mismatch can be described and how to extract model parameters. This work was done using physically based models for the drain current. However, the physics behind the variability itself was not investigated. In this chapter we will go one level of abstraction deeper and investigate the *physical content* of MOSFET mismatch. By physical content we mean *the origin of the microscopic differences between two transistors* and *how these microscopic differences affect macroscopic transistor operation.* Knowledge about the physical origins of MOSFET mismatch allows the refinement of models, provides information about the dominant mismatch causing mechanisms, and can ultimately lead to technologies with a better matching performance.

In literature, mainly the physical origins of threshold voltage mismatch are examined. They are found to be related to doping fluctuations, fluctuations in gate depletion and fluctuations in boron penetration [3, 21]. In order to calculate their impacts, generally, charge sheet modeling is applied [5, 10, 11, 16, 19, 30, 32, 37, 38, 51, 65, 75, 76] or 2D or 3D device simulations are performed [11, 12, 77–94]. For short-channel devices two-dimensional field-effects influence device behavior, which to first order are modeled in [17, 22, 31, 37, 38, 51, 61, 75, 88, 92, 95]. Finally, by means of simulations the impact of quantum mechanics has been examined [79, 83, 84, 90].

Although much work is ongoing, a complete understanding of the physical origins of mismatch is still lacking. For instance, doping fluctuations only manage to explain about half of the experimentally observed mismatch in the threshold voltage. In this chapter theories will be presented that model the impact of doping fluctuations in the channel, doping

fluctuations in the gate, fluctuations in oxide charge, and fluctuations in surface roughness. As an introduction, we will start in section 4.1 by deriving the basic equations of MOSFET operation. The second section again derives these equations, but now in the presence of microscopic fluctuations. It will be found that the generally applied charge-sheet approach is only valid in strong inversions at low values of the drain bias. New models will be provided for the weak inversion regime, and for the strong inversion regime at higher values of the drain bias. Furthermore, short- and narrow-channel effects and the device symmetry are examined. The actual calculation of the impact of the above mentioned physical effects takes place in section 4.3. Quantum-mechanical and mobility effects will be taken into account. We combine all our models to compare them to experimental data. They are extensively tested by investigating the bias dependencies of the mismatch. With our theories we will manage to understand most of the experimental results. Section 4.4 concludes this chapter.

4.1 Basic operation of the MOS transistor

This section gives a brief overview of the basic operation of the MOS transistor. For more extensive descriptions we refer to [96, 97]. This section is organized as follows: In the first subsection the regions of operation of the MOSFET are introduced, and current expressions are derived for long-channel transistors. The second subsection discusses short- and narrow-channel effects and the impact of halos. Subsections 4.1.3 and 4.1.4 describe the impact of gate depletion and quantummechanical effects, respectively. Subsection 4.1.5 looks deeper into mobility determining effects.

4.1.1 Regions of operation and current expressions

Figure 4.1 shows the basic structure of the n-type MOSFET. Depending on the technology node, for modern-day devices, doping levels range from $10^{17} - 5 \cdot 10^{18}$ cm^{-3} [1] in the channel region and $5{\cdot}10^{19} - 10^{21}$ cm^{-3} [2] in the extension regions and in the poly-silicon gate. When the voltage applied between the gate and bulk (V_{GB}) is equal to their difference in work function (ϕ_{MS}), the energy bands show no bending. This situation is called the flat-band condition and is displayed in

[1] The doping type is boron (or possibly Indium) for nMOS transistors and arsenic for pMOS transistors.

[2] The doping type is arsenic for nMOS transistors and boron (or possibly Indium) for pMOS transistors.

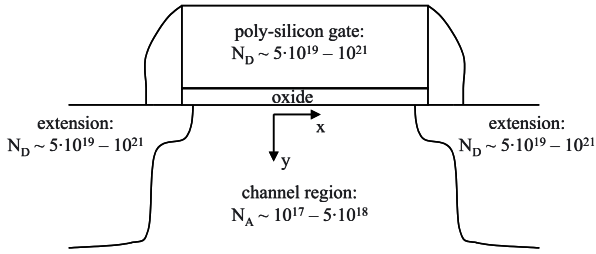


Figure 4.1. Schematic drawing of the MOSFET

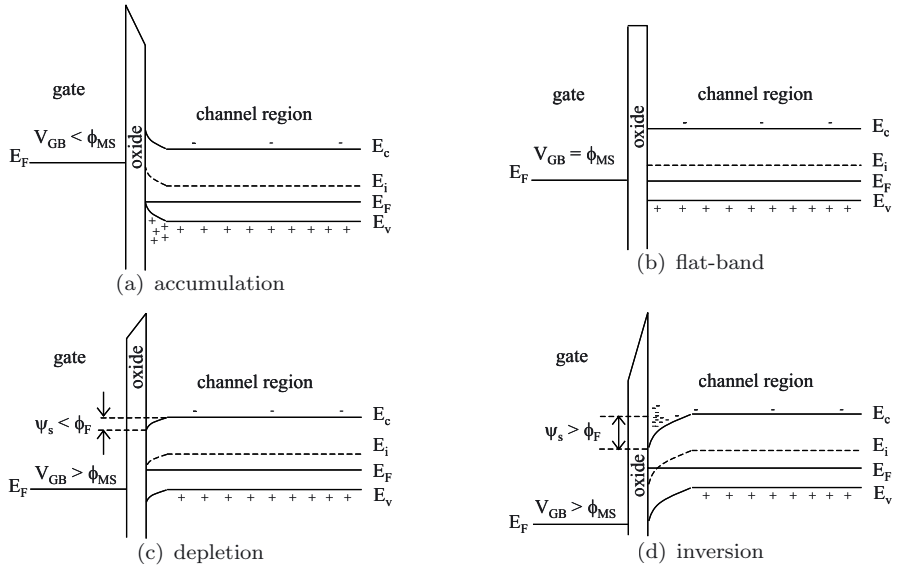


Figure 4.2. Schematic band diagrams for the four regions of MOSFET operation

figure 4.2b. When the gate-to-bulk bias is smaller, holes accumulate to the surface (figure 4.2a). For larger V_{GB} holes are pushed away from the oxide-silicon interface and a depletion layer appears (figure 4.2c). When V_{GB} is increased further, the intrinsic energy (E_i) at the surface will become smaller than the Fermi level (E_F), which causes the concentration of surface electrons (n) to become larger than the concentration of holes (figure 4.2d). This situation is called inversion and the layer of electrons at the interface is called the inversion layer. In most cases, the MOSFET operates in this regime. The difference between the intrinsic energy in the bulk and the Fermi energy is given by:

$$E_i - E_F = kT \ln\left(\frac{N_A}{n_i}\right), \tag{4.1}$$

where k is Boltzmann's constant, N_A the channel-doping concentration, n_i the intrinsic carrier concentration and T the temperature. In other words, the transistor operates in inversion when the V_{GB} causes $\psi_s > \phi_F$. The Fermi potential $\phi_F = (E_i - E_F)/q$ and ψ_s is the surface potential. When V_{GB} is increased, at first, the depletion region under the gate will widen, the surface potential (ψ_s) will increase, but the electron concentration at the interface remains low. This is called weak inversion. However, when the surface potential reaches $\psi_s = 2\phi_F$, it becomes energetically more favorable to add electrons to the surface than to increase the width of the depletion layer. Therefore, the electron concentration in the inversion layer becomes significant, while the depletion layer width is constant and the surface potential remains fixed. This is called strong inversion.

We will proceed by deriving the current expressions for the inversion regimes. In general the electron current-density (J_n)[3] can be expressed as:

$$J_n = -qn\mu_n \frac{d\psi_s}{dx} + kT\mu_n \frac{dn}{dx}, \tag{4.2}$$

where μ_n is the electron mobility. The first term on the right-hand side represents the drift-current component and the second term represents the diffusion current. In order to derive current expressions it will first be assumed that the source and drain are at the same potential (V_C). We need to calculate: 1) How the electron concentration in the inversion layer depends on the surface potential and V_C, 2) how the surface potential depends on V_{GC} and V_{CB}, and finally 3) how, for unequal source and drain potential, the surface potential and electron concentration vary laterally.

From Boltzmann statistics it follows that:

$$n(x) = n_i e^{(\psi_s(x) - (\phi_F + V_{CB}))q/kT} = N_A e^{(\psi_s(x) - (2\phi_F + V_{CB}))q/kT}. \tag{4.3}$$

With the bulk taken as reference, V_{CB} equals the increase in Fermi potential due to the contact with the source and drain. In order to find the potential and charge distribution, Poisson's equation needs to be solved:

$$\frac{d^2\psi}{dx^2} = -\frac{q}{\epsilon_{si}} \left(p(y) - n(y) - N_A(y)\right) \tag{4.4}$$

For the total charge under the gate (Q_s) in inversion it can be found that these equations yield:

$$Q_s \cong Q_D + Q_i \cong -\sqrt{2\epsilon_{si} q N_A \left(\psi_s + (kT/q) e^{(\psi_s - (2\phi_F + V_{CB}))q/kT}\right)}, \tag{4.5}$$

[3] In n-type transistors the hole current-density is negligible (and vice versa).

where Q_D is the depletion layer charge and Q_i is the inversion layer charge. It follows that if Q_D is known, Q_i is also known. To calculate Q_D, the depletion approximation is used: In the depletion region ($0 < y < W_D$) the hole concentration $p = 0$, while outside this region $p = N_A$ ($y > W_D$). From (4.4) it follows that a charge sheet at position y with charge $qN_A(y)dy$ gives rise to a field ($E_{N_A(y)}$) of:

$$E_{N_A(y)}(y') = \begin{cases} 0 & y' > y \\ qN_A(y)dy/\epsilon_{si} & 0 < y' < y \\ qN_A(y)dy/\epsilon_{ox} & -t_{ox} < y' < 0 \end{cases} . \tag{4.6}$$

This gives for the surface potential:

$$\psi_s = \frac{q}{\epsilon_{si}} \int_0^{W_D} yN_A(y)dy, \tag{4.7}$$

where W_D is the thickness of the depletion layer. It is assumed that Q_i is fully located at $y = 0$. For a uniform doping concentration ($N_A(y) = N_A$) it follows that:

$$W_D = \sqrt{\frac{2\epsilon_{si}\psi_s}{qN_A}}, \tag{4.8}$$

$$Q_D = -\sqrt{2\epsilon_{si}qN_A\psi_s}, \tag{4.9}$$

In weak inversion the exponential term in (4.5) is much smaller than ψ_s. Using a first order Taylor expansion and combining (4.5) and (4.9) gives for the inversion-layer charge:

$$Q_i \cong -\sqrt{\frac{\epsilon_{si}qN_A}{2\psi_s}} \frac{kT}{q} e^{(\psi_s - (2\phi_F + V_{CB}))q/kT}. \tag{4.10}$$

In order to relate ψ_s to V_{GC} and V_{CB}, the following potential balance is used:

$$V_{GC} + V_{CB} = \phi_{MS} + \psi_s + V_{ox}. \tag{4.11}$$

The potential over the oxide (V_{ox}) is equal to:

$$V_{ox} = \frac{qt_{ox}}{\epsilon_{ox}} \int_0^{W_D} N_A(y)dy, \tag{4.12}$$

which for a uniform doping concentration results in:

$$V_{GB} = \phi_{MS} + \psi_s + \frac{t_{ox}\sqrt{2\epsilon_{si}qN_A\psi_s}}{\epsilon_{ox}}. \tag{4.13}$$

In strong inversion the surface potential can be assumed fixed at $\psi_s = 2\phi_F + V_{CB} \equiv \phi_B + V_{CB}$, as was mentioned before. The lower V_{GC} boundary of this regime is called the threshold voltage (V_T). From (4.11) it is seen to be equal to:

$$V_T = \phi_{MS} + \phi_B + \frac{t_{ox}\sqrt{2\epsilon_{si}qN_A(\phi_B + V_{CB})}}{\epsilon_{ox}}. \tag{4.14}$$

When $V_{GC} > V_T$, an increase in V_{GC} mainly results in an increase of the inversion layer charge. In other words, the device behaves as a capacitor over which a voltage $V_{GC} - V_T$ is applied. For the inversion-layer charge this yields:

$$Q_i = -\frac{\epsilon_{ox}}{t_{ox}}(V_{GC} - V_T). \tag{4.15}$$

The approximation of $\psi_s = 2\phi_F + V_{CB}$, independent of V_{GC}, is checked by introducing this equation in (4.5). Since now the exponential term is dominant it follows that:

$$\phi_B = 2\phi_F + \frac{2kT}{q}\ln\left(\frac{\epsilon_{ox}(V_{GC} - V_T)}{t_{ox}\sqrt{2kT\epsilon_{si}N_A}} + \sqrt{\frac{q(2\phi_F + V_{CB})}{kT}}\right) \cong 2\phi_F + \frac{5kT}{q}, \tag{4.16}$$

which is a bit more accurate than putting $\phi_B = 2\phi_F$.

We will proceed with calculating the current as a function of V_{GS}, V_{DS} and V_{BS}. The gradual channel approximation is used: The lateral change in potential is small enough, so that locally the structure is described by the equations that were derived earlier in this subsection. However, V_{GC} and V_{CB} now are functions of the lateral position (x). At $x = 0$, $V_{GC} = V_{GS}$ and $V_{CB} = -V_{BS}$. At $x = L$, $V_{GC} = V_{GS} - V_{DS}$ and $V_{CB} = V_{DS} - V_{BS}$.

In weak inversion the current is caused by diffusion. It follows from (4.2) and (4.10) that:

$$I_D = kT\mu_n\frac{W}{L}\left(Q_i|_{x=0} - Q_i|_{x=L}\right) = \tag{4.17}$$

$$= \frac{WI_0}{L}e^{q(\psi_s - 2\phi_F + V_{BS})/kT}\left(1 - e^{-qV_{DS}/kT}\right),$$

$$I_0 = \mu_n\sqrt{\frac{\epsilon_{si}q^3N_A}{2\psi_s}}\left(\frac{kT}{q}\right)^2.$$

The surface potential is calculated with (4.13). Note that this potential is independent of the lateral position, in accordance with neglecting the

drift component. Often, (4.13) is approximated by:

$$\psi_s + V_{BS} \cong \frac{V_{GS} - V_T}{1 + \delta}, \tag{4.18}$$

where $1/(1+\delta)$ to first order models the sensitivity of V_{CS} to V_{GS} and:

$$\delta = \frac{t_{ox}}{\epsilon_{ox}} \sqrt{\frac{\epsilon_{si} q N_A}{2(\phi_B - V_{BS})}} = \frac{\epsilon_{si} t_{ox}}{\epsilon_{ox} W_D} = n - 1. \tag{4.19}$$

In strong inversion, the drift component in (4.2) dominates. Combining this with (4.15) and realizing that $\psi_s = \phi_B + V_{CB}$ gives:

$$I_D = \frac{W \epsilon_{ox}}{t_{ox}} (V_{GS} - V_T - V_{CS}) \frac{dV_{CS}}{dx} \cong \tag{4.20}$$

$$\cong \frac{W \epsilon_{ox}}{t_{ox}} (V_{GS} - V_T|_{V_{CB} = -V_{BS}} - (1+\delta) V_{CS}) \frac{dV_{CS}}{dx},$$

where $\delta \cdot V_{CS}$ models, to first order, the V_{CS} dependence of the threshold voltage. From now on, unless explicitly mentioned, the symbol V_T is used for the threshold voltage at the source side of the transistor. Solving (4.20) gives:

$$V_{CS} = \frac{V_{GS} - V_T - \sqrt{(V_{GS} - V_T)^2 - 2(1+\delta)(I_D/\beta)(x/L) + C}}{1 + \delta}, \tag{4.21}$$

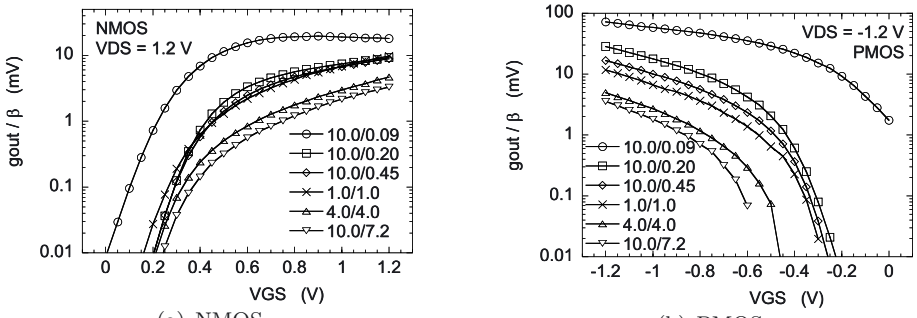
where the current factor $\beta = W \mu_n \epsilon_{ox} / L t_{ox}$ and the constant $C = 0$, as follows from the boundary condition $V_{CS}(x = 0) = 0$. The drain current is solved from the other boundary condition $V_{CS}(x = L) = V_{DS}$, which gives:

$$I_D = \beta (V_{GS} - V_T - (1+\delta) V_{DS}/2) V_{DS}. \tag{4.22}$$

This equation is valid in the linear regime. At high drain bias the inversion layer is pinched of at the drain side. In (4.22) V_{DS} has to be replaced by the saturation voltage (V_{DSsat}), which is calculated by equating $dI_D/dV_{DS} = g_{out}$, which results in:

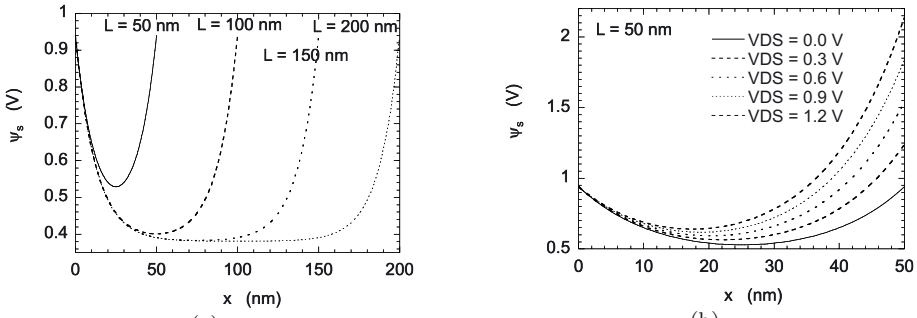
$$V_{DSsat} = \frac{V_{GS} - V_T - g_{out}/\beta}{1 + \delta}, \tag{4.23}$$

When $V_{DS} > V_{DSsat}$ the transistor is called to be operating in saturation. Figure 4.3 plots the output conductance over the current factor (g_{out}/β) as a function of the gate bias for a 130 nm technology. It is seen that g_{out}/β affects the saturation voltage only for short transistors. Note that (4.22) and (4.23) differ somewhat from (2.17), (2.24) and (3.1). In the previous chapters it was assumed that $\delta = 0$ and $g_{out} = 0$, while in calculating (4.23) the drain-bias dependence of the mobility was not included.



(a) NMOS (b) PMOS

Figure 4.3. Output conductance over the current factor as a function of the gate bias for a 130 nm technology with a nominal gate length of 90 nm, $|V_{DD}| = 1.2$ V and $t_{ox} = 1.5$ nm



(a) (b)

Figure 4.4. Surface potential (ψ_s) as a function of the lateral position (x) with the length (a) or drain bias (b) as a parameter. $N_A = 1 \cdot 10^{18}$ cm^{-3}, $t_{ox} = 2.2$ nm.

4.1.2 Short- and narrow-channel effects

In the previous subsection it was assumed that the channel length of the transistor is long. Close to the source and drain junctions, one has to use the two dimensional Poisson equation to find the electric field. For long-channel devices these 2D effects are negligible. However, for short devices they start to play a role. This subsection discusses this short-channel effect (SCE) and channel-length modulation, which is another SCE. Besides SCEs, also the narrow-channel effect will be discussed.

2D field effects. Figure 4.4a shows the surface potential in weak inversion as a function of the lateral position for several transistor lengths at $V_{DS} = 0$ V. Figure 4.4b shows the surface potential for a short device, as a function of the lateral position, for several values of the drain bias.

The surface potential can be approximated by [98]:

$$\psi_s(x) \approx \psi_{si}^0 + (\psi_{siend} - \psi_{si}^0)\frac{\sinh((x - x_{ibegin})/\lambda_i)}{\sinh(L_i/\lambda_i)} + \tag{4.24}$$

$$+ (\psi_{sibegin} - \psi_{si}^0)\frac{\sinh((x_{iend} - x)/\lambda_i)}{\sinh(L_i/\lambda_i)},$$

where $x_{begin} = 0$, $x_{end} = L$, $\psi_{begin} = \psi_s(x = 0) = \psi_{bi} - V_{BS}$, $\psi_{end} = \psi_s(x = L) = \psi_{bi} - V_{BS} + V_{DS}$, the built-in potential $\psi_{bi} = E_g/2q + \phi_B$ and ψ_s^0 is the long-channel surface potential, given by (4.11). The meaning of the subscript i will become clear when halos are introduced later in this subsection. The parameter λ models the rate of change of the surface potential with the lateral position at the source and drain end of the transistor. It is given by:

$$\lambda = \xi\sqrt{\frac{\epsilon_{si}}{\epsilon_{ox}} t_{ox} W_D}, \tag{4.25}$$

where ξ is equal to 1 for abrupt junctions, but is generally used as a fitting parameter.

Two effects are observed. Firstly, for short transistors it is seen that the potential barrier between source and drain is smaller than $2\phi_F$, which also causes a smaller threshold voltage. The decrease in barrier height is calculated at the location of minimal potential:

$$x_{min} \cong \frac{L}{2} - \frac{\lambda}{2}\ln\left(\frac{\psi_{end} - \psi_s^0}{\psi_{begin} - \psi_s^0}\right), \tag{4.26}$$

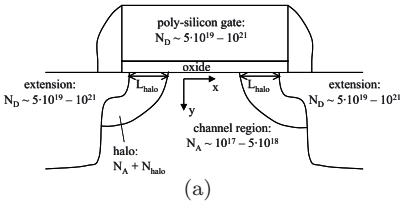
$$\Delta\psi_s = \psi_s(x_{min}) - \psi_s^0 \cong 2\sqrt{(\psi_{end} - \psi_s^0)(\psi_{begin} - \psi_s^0)}e^{-L/2\lambda} \tag{4.27}$$

which has been derived, using the approximation $\sinh(z) \cong e^z/2$. The difference is seen to increase with increasing drain bias, which is called drain induced barrier lowering (DIBL). It is also observed that $d\psi_s(x_{min})/d\psi_s^0 < 1$. This explains the increase in subthreshold swing and decrease in dV_T/dV_{BS} for short devices, since both the gate- and bulk-bias dependencies of ψ_s are mainly determined by $\psi_s^0(V_{GS}, V_{BS})$.

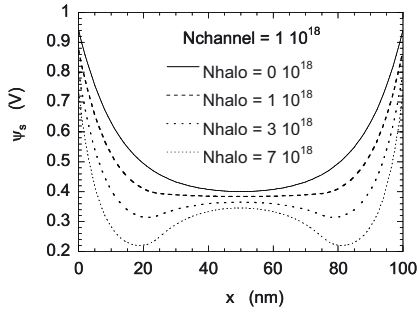
Secondly, a significant part of the channel is needed to build up the potential. This results in a shorter effective channel length (L_{eff}) than the metallurgical channel length (L_{met}). As measure of the effective channel length we take[4]:

$$\text{weak inversion:} \qquad L_{eff} = \frac{\int_0^L Q_i(\psi_s(x))dx}{Q_i(\psi_s(x_{min}))}, \tag{4.28}$$

[4]At this point it is useful to briefly summarize the different lengths that are used in this book. The gate length (L_{gate}) refers to the length of the poly-silicon gate. The gate length



(a)



(b)

Figure 4.5. a) Schematic drawing of a MOSFET with halos. b) Surface potential (ψ_s) as a function of the lateral position (x) with the halo dose as a parameter. $L = 100$ nm, $L_{halo} = 25$ nm and $t_{ox} = 2.2$ nm.

$$\text{strong inversion:} \quad L_{eff} = L_{met} - \lambda_{source} - \lambda_{drain}. \tag{4.29}$$

The inversion-layer charge in weak inversion is calculated with (4.10). Further note that it follows from (4.25) that for unequal source and drain bias $\lambda_{source} \neq \lambda_{drain}$, since the depletion layer widths at both sides differ.

We will now make the following approximation: The short-channel transistor is assumed to behave as a long-channel transistor with $L = L_{eff}$ and $\psi_s = \psi_s^0 + \Delta\psi_s$. The effective channel length replaces L in the formula for the current factor. The increase in surface potential lowers the threshold voltage by:

$$\Delta V_T(L) \equiv V_T(L) - V_{Tlw} \cong -(1+\delta)\Delta\psi_s, \tag{4.30}$$

where V_{Tlw} is the long-channel threshold voltage.

To counter the short-channel effect, in modern-day devices extra doping is implanted around the source and drain regions (see figure 4.5a). These regions with a higher doping level are called halos. Figure 4.5b shows the surface potential as a function of the lateral position for transistors with different halo doses (N_{halo}). The transistor is now divided in three regions: 1) halo at source side, 2) center, 3) halo at drain side. In each of these three regions (4.24) is valid. The subscript i denotes the region. The boundary conditions are $\psi_{1begin} = \psi_{bi} - V_{BS}$, $\lim_{x\uparrow x_{1end}} d\psi_s/dx =$

of the minimum sized digital transistor is called the nominal gate length ($L_{nominal}$). The metallurgical channel length (L_{met} or $L_{channel}$) is equal to the distance between the source and drain. The effective channel length (L_{eff}) is equal to the electrical channel length. In this book, the length L usually refers to the metallurgical channel length. For long-channel transistors the differences between these lengths are negligible.

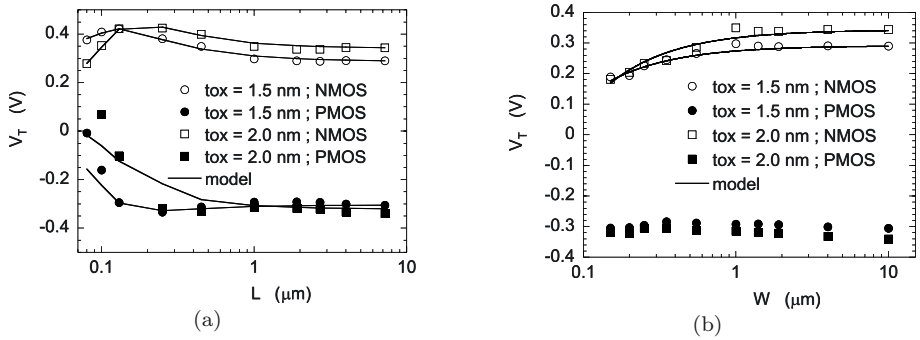


Figure 4.6. Threshold voltage as a function of the gate length (a) and gate width (b) for two 130 nm technologies. The first has a nominal gate length of 90 nm with $t_{ox} = 1.5$ nm and $|V_{DD}| = 1.2$ V. The second has a nominal gate length of 130 nm with $t_{ox} = 2.0$ nm and $|V_{DD}| = 1.5$ V. Model parameters are listed in table 4.1. $V_{DS} = 50$ mV.

$\lim_{x\downarrow x_{2begin}} d\psi_s/dx$, $\lim_{x\uparrow x_{2end}} d\psi_s/dx = \lim_{x\downarrow x_{3begin}} d\psi_s/dx$ and $\psi_{3end} = \psi_{bi} - V_{BS} + V_{DS}$. The overall shift in surface potential is approximated by:

$$\Delta\psi_s = \frac{1}{L_{eff}} \int_{\lambda_{source}}^{L-\lambda_{drain}} \psi_s(x)dx - \psi^0_{s2,long} \tag{4.31}$$

The subscript *long* is added to ψ^0_{s2} to distinguish it from the short-device case, in which the doping level in region 2 is determined by overlapping halos. The threshold voltage is again calculated with (4.30).
Figure 4.6a compares experimental values of the threshold voltage with calculated values. Model parameters are presented in table 4.1. Results for the NMOS transistors are seen to be well described. The PMOS devices, that suffer more strongly from the short channel effect, are less well described. This can be explained by the usage of (4.8) to calculate the depletion layer width. For too strong 2D field-effects the depletion layer width increases and this approximation is not valid.

Channel-length modulation. In strong inversion, when the device is operated in saturation ($V_{DS} > V_{DSsat}$), the channel is pinched off when the potential reaches $V(x) = V_{DSsat}$. This point is not exactly located at $x = L$, but a distance l_p closer to the source. This distance can be approximated by [96]:

$$l_p = \sqrt{\frac{2\epsilon_{si}}{qN_A}} \left(\sqrt{\frac{\epsilon_{si}E_1^2}{2qN_A} + (V_{DS} - V_{DSsat})} - \sqrt{\frac{\epsilon_{si}E_1^2}{2qN_A}} \right), \tag{4.32}$$

Table 4.1. Model parameters describing the short- and narrow-channel effects related to V_T, $\sigma^2_{\Delta V_T}$ and $\sigma^2_{\Delta\beta/\beta}$

t_{ox} (nm)	1.5	1.5	2.0	2.0
$L_{nominal}$ (nm)	90	90	130	130
type	NMOS	PMOS	NMOS	PMOS
V_{DD} (V)	1.2	-1.2	1.5	-1.5
N_A (cm^{-3})	$6 \cdot 10^{17}$	$6 \cdot 10^{17}$	$5 \cdot 10^{17}$	$5 \cdot 10^{17}$
V_{Tlong} (V)	0.291	-0.296	0.343	-0.325
$A_{0,\Delta V_T}$ (mVμm)	3.7	2.3	3.8	2.4
$A_{0,\Delta\beta/\beta}$ (%μm)	1.17	0.86	1.00	0.86
L_{halo} (nm)	135	130	130	130
N_{halo} (cm^{-3})	$7.5 \cdot 10^{17}$	$6.0 \cdot 10^{17}$	$9.5 \cdot 10^{17}$	$6.0 \cdot 10^{17}$
ξ (-)	1.0	2.3	2.5	4.5
$\Delta L_{\Delta V_T}$ (nm)	40	150	65	155
$\Delta L_{\Delta\beta/\beta}$ (nm)	60	100	80	155
W_{narrow} (nm)	50	50	70	50
$V_{Tnarrow}$ (V)	0.118	-0.296	0.156	-0.325
$A_{\Delta V_T,narrow}$ (mVμm)	1.65	3.55	2.45	3.55
$A_{\Delta beta/\beta,narrow}$ (%μm)	1.17	1.6	1.17	1.6

where $E_1 = d\psi_s/dx|_{x=L-l_p} \sim 10^4 - 2 \cdot 10^5$ V/cm. To properly describe the current, in (4.22) and (4.23), L needs to be replaced by $L - l_p$. This effect is called channel-length modulation.

Narrow-channel effect. For narrow transistors deviations from wide-channel behavior can be expected. For transistors with shallow-trench isolation (STI), the gate curves a bit around the edge. This causes a larger effective gate-area at the side of the channel, which results in a lowering of the threshold voltage. For NMOS transistors, close to the isolation the boron channel-doping can be reduced due to segregation of dopants into the STI. This also results in a lower threshold voltage. Finally, stress induced by the isolation can also affect transistor operation through a change in the band gap and a change in the mobility.
These effects can be modeled by dividing the transistor in three parallel segments: The center segment with width $W_{middle} = W - 2W_{narrow}$ has the 'normal' threshold voltage (V_{Tlw}), while the two transistors at the edge with width W_{narrow} have a threshold voltage ($V_{Tnarrow}$) adjusted for the narrow-channel effects. The overall threshold voltage (V_T) is

given by:

$$V_T(W) = \frac{W_{middle}}{W} V_{Tlw} + \frac{2W_{narrow}}{W} V_{Tnarrow}. \tag{4.33}$$

Figure 4.6b shows that this model gives a good description of the threshold voltage as a function of the width. Model parameters are listed in table 4.1.

4.1.3 Gate depletion

The total amount of charge in the MOSFET is equal to 0. Therefore, the negative charge under the oxide is equalled by the same amount of positive charge in the gate. This results in a small depletion layer on top of the oxide, that decreases the total gate-to-channel capacitance (C_{GC}). The equivalent increase in oxide thickness (t_{GD}) is defined as $t_{GD} \equiv \epsilon_{ox}/C_{GC} - t_{ox}$. Using the depletion approximation[5], we can write:

$$t_{GD} = -\frac{\epsilon_{ox} Q_s}{\epsilon_{si} q N_p} \cong \frac{\epsilon_{ox}^2 (V_{GS} - \phi_{MS} - \phi_B)}{\epsilon_{si}(t_{ox} + t_{GD}) q N_p} = \tag{4.34}$$

$$= \sqrt{\left(\frac{t_{ox}}{2}\right)^2 + \frac{\epsilon_{ox}^2 (V_{GS} - \phi_{MS} - \phi_B)}{\epsilon_{si} q N_p}} - \frac{t_{ox}}{2},$$

where N_p is the doping concentration in the poly gate at the interface with the oxide. This concentration can be significantly lower than the average doping concentration in the gate and is estimated to have an approximate value of $N_p \sim 5 \cdot 10^{19}$ cm^{-3}. At an oxide thickness of $t_{ox} = 1.5$ nm and gate bias of $V_{GS} = 1.2$ V[6] this results in an effective increase in oxide thickness of $t_{GD} = 0.75$ nm.
In the calculation of the drain current t_{ox} needs to be replaced by $(t_{ox} + t_{GD})$. Note that the drain-current is decreased, because of the decrease in oxide capacitance directly (see (4.15)) and because of the related increase in threshold voltage (see (4.14)).

4.1.4 Quantummechanical effects

Figure 4.7 schematically plots the potential in the channel region as a function of the distance from the oxide-silicon interface. It is seen that at the surface the potential can be approximated by a triangular

[5]The depletion-layer width in the gate ($(\epsilon_{si}/\epsilon_{ox})t_{GD}$) is of the same order of magnitude as the Debye-length (compensated for Fermi-Dirac statistics). Strictly speaking, this means that the use of the depletion approximation is not valid.
[6]$\phi_{MS} + \phi_B \sim 0$.

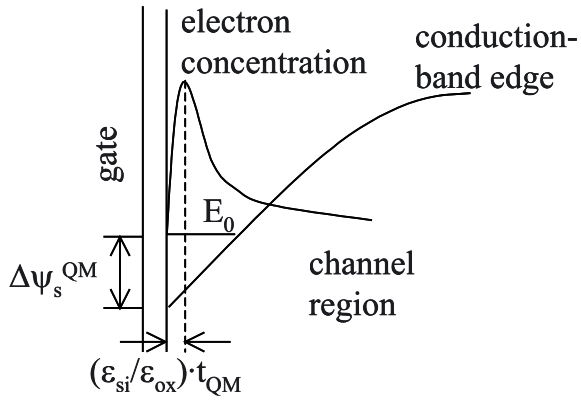


Figure 4.7. Schematic drawing of the potential and electron concentration taking quantummechanical effects into account

well in which the electron wave-functions are confined. Also the electron distribution of the ground state is plotted in figure 4.7. Two effects are observed. Firstly, the ground-state has an energy, that is slightly higher than ψ_s and extra band bending ($\Delta\psi_S^{QM}$) is required to reach the threshold condition [97]:

$$\Delta\psi_s^{QM} \approx B_{QM1} \cdot E_s^{2/3} - \frac{kT}{q}\ln\left(\frac{E_s}{E_{QM2}}\right), \tag{4.35}$$

where the surface field $E_s = Q_D/\epsilon_{si}$, $B_{QM1} = 1.73 \cdot 10^{-5}$ $\mathrm{V}^{1/3}\mathrm{cm}^{2/3}$ and $E_{QM2} = 2.02 \cdot 10^5$ Vcm^{-1}. This approximation only takes the lowest energy subband into account, which is accurate when $E_s \gtrsim 5{\cdot}10^5$ Vcm^{-1} or $N_A \gtrsim 1 \cdot 10^{18}$ cm^{-3}.
Secondly, it is observed that the peak electron concentration is not located at the interface but a certain distance $(\epsilon_{si}/\epsilon_{ox})t_{QM}$ away from it:

$$\frac{\epsilon_{si}t_{QM}}{\epsilon_{ox}} = \frac{B_{QM3}}{(Q_D + \frac{11}{32}Q_i)^{1/3}}, \tag{4.36}$$

where $B_{QM3} = 1.25 \cdot 10^{-9}$ $\mathrm{cm}^{1/3}\mathrm{C}^{-1/3}$. This results in an increase of the effective oxide thickness of $t_{QM} \sim 0.4$ nm, that lowers the current factor.

4.1.5 Low field mobility

As was mentioned in section 2.3.3, the mobility is determined by several scattering mechanisms. It can be split up in bulk mobility (μ_B), surface and fixed oxide-charge scattering (μ_{fc}), Coulomb scattering (μ_C)

and surface roughness scattering μ_{sr}. The overall mobility (μ) is calculated by Matthiessen's rule:

$$\frac{1}{\mu} = \frac{1}{\mu_B} + \frac{1}{\mu_{fc}} + \frac{1}{\mu_C} + \frac{1}{\mu_{sr}}. \tag{4.37}$$

The mobilities will be expressed in terms of the effective field (E_{eff}):

$$E_{eff} = |Q_B + \eta Q_i|/\epsilon_{si}. \tag{4.38}$$

The parameter η is related to the inversion layer thickness. Theoretically it is equal to 11/32, when only the lowest subband is taken into account. Experimentally it is found that $\eta = 1/2$ for electrons and $\eta = 1/3$ for holes. The value of 1/2 for electrons is due to the occupancy of higher subbands. The value of 1/3 for holes is very close to the theoretical 11/32.

A physically correct approach to model the different components is presented in e.g. [99–101]. The full geometry of the problem needs to be taken into account, e.g. by using Green's functions. Generally, mobility is calculated by averaging out over all possible device configurations. Fluctuations could be introduced by realizing that a device with finite dimensions can, in itself, not possess all possible configurations. The resulting expressions are quite complicated. Therefore, we choose to use the simpler semi-empirical expressions published in [102] and listed below:

$$\mu_{fc} = \frac{z_\mu}{3.2 \cdot 10^{-9} p_\mu}, \tag{4.39}$$

$$p_\mu = 0.09 + 9.06 \cdot 10^{-13} (z_\mu/|Q_i|)^{1/4} N_f, \tag{4.40}$$

$$z_\mu = 0.388/E_{eff} + 1.73 \cdot 10^{-5}/E_{eff}^{1/3}, \tag{4.41}$$

$$\mu_C = \frac{1.1 \cdot 10^{21}}{\ln(1+\gamma_{BH}^2) - \frac{\gamma_{BH}^2}{1+\gamma_{BH}^2}} \frac{m_C}{N_A}, \tag{4.42}$$

$$\gamma_{BH}^2 = 3.2 z_\mu/|Q_i|, \tag{4.43}$$

$$\mu_{sr} = K_{sr}/E_{eff}^2, \tag{4.44}$$

where K_{sr} and m_C are a proportionality constants and N_f is the fixed oxide-charge density.

Figure 4.8a shows a fit of the model to experimental transconductance curves as a function of the gate bias for several values of the bulk bias.

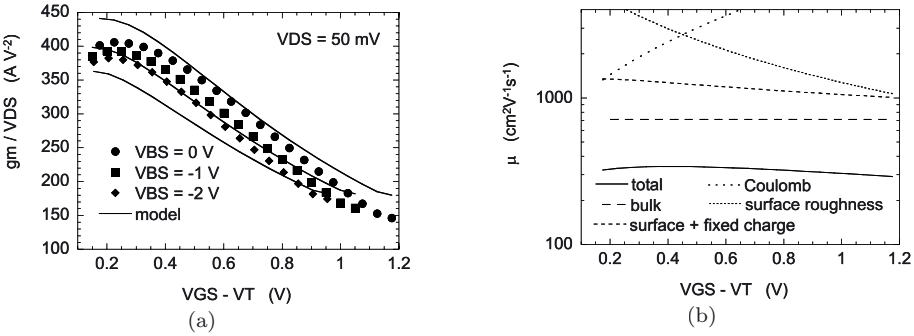


Figure 4.8. a) Experimental and modeled transconductance as a function of the gate overdrive at three values of the bulk bias. b) Total mobility and its components as a function of the gate overdrive at $V_{BS} = 0$ V. The experimental data was obtained from NMOS transistors with $t_{ox} = 2$ nm.

The experimental curves seem to be reasonably well described. The results of the fit are $K_{sr} = 1.0{\cdot}10^{15}$ Vs^{-1}, $m_C = 0.70$, $N_f = 1.16{\cdot}10^{11}$ cm^{-2} and $\mu_B = 715$ cm^{2}V^{-1}s^{-1}. The large amount of fixed oxide-charge is typical for heavily nitrided gate oxides. The doping concentration close to the oxide-silicon interface was obtained by SIMS and is equal to $N_A = 3.1 \cdot 10^{17}$ cm^{-3}.
Figure 4.8b shows the overall mobility and the magnitude of its components. The main components are μ_B and μ_{fc}. At low gate bias, Coulomb scattering starts to play a role, but at high gate bias the inversion layer screens the dopants. In this region surface-roughness scattering becomes important.

4.2 Mismatch in the drain current

We will now proceed by examining the impact of a mismatch in the threshold voltage and of a mismatch in the current factor on the drain current. In subsection 2.4.1 and literature it is assumed that the overall mismatch in a parameter (ΔP) can be calculated by averaging out the microscopic mismatch ($\delta P(x, z)$) over the area of the transistor. The impact on the drain current follows from (2.3). This is called the charge-sheet approximation. This approximation is examined by calculating the mismatch in the drain current directly from the microscopic mismatch, using (4.2). We find that the charge-sheet approximation is only valid in strong inversion at low values of the drain bias. Deviations for long-channel devices are calculated in subsection 4.2.1 for the weak inversion regime and in subsection 4.2.2 for the strong inversion regime. Subsection 4.2.3 discusses short- and narrow-channel effects for which models

from literature are applied. Subsection 4.2.4 investigates the differences between the weak and strong inversion regimes, that were earlier observed in chapter 2 and chapter 3. By closely examining the averaging effects, we are able to explain most of the differences. Finally, in subsection 4.2.5, the symmetry of the MOSFET is examined, which is closely related to its matching properties.

4.2.1 Solution of the current equation in weak inversion

In weak inversion, the mismatch in drain current will mainly be caused by fluctuations in the surface potential $\psi_s = \psi_{s0} + \delta\psi_s(x)$. These are expected to be dominating, since ψ_s is the only fluctuating parameter in the exponential of (4.17). When we write $n(x) = \delta f_n(x) \cdot n_0(x)$ and $\delta f_n(x) = e^{q\delta\psi_s(x)/kT}$, it follows from (4.2) and (4.3) that:

$$J_n = \frac{1}{L}\int_0^L \left(-q\delta f_n(x) n_0(x)\mu_n \frac{\delta\psi_s(x)}{dx} + kT\mu_n \frac{d\delta f_n(x)}{dx} n_0(x)\right) dx + \tag{4.45}$$

$$+\frac{1}{L}\int_0^L \left(kT\mu_n \delta f_n(x)\frac{dn_0(x)}{dx}\right) dx = \frac{1}{L}\int_0^L \left(kT\mu_n \delta f_n(x)\frac{dn_0(x)}{dx}\right) dx.$$

This equation has the same shape as:

$$J_n = \frac{1}{\rho(x)}\frac{dV}{dx}. \tag{4.46}$$

In other words, the resistivity that the current locally experiences is proportional to $1/\delta f_n(x)$ and the driving force is proportional to the concentration gradient. In the two dimensional case $dn_0(x)/dx$ does not vary with x, i.e. a microscopic difference at the source side of the transistor has the same impact as a microscopic difference in the middle or at the drain side[7]. In order to test this theory 2D simulations were performed in MEDICI [104] of a MOSFET which has a slightly higher doping concentration between $x_h - 5.5$ nm and $x_h + 5.5$ nm (see figure 4.9a)[8]. To avoid errors due to small differences in grid, it was made sure that compared simulations had exactly the same grid. Figure 4.10a shows that the relative decrease in drain current due to the extra doping is independent of x_h and of the drain bias. It is also observed that the theory out of subsection 4.1.1 gives a good prediction of the shift.

[7] This conclusion was also reached in [103]

[8] At low drain bias, a similar kind of analysis was performed in [75]

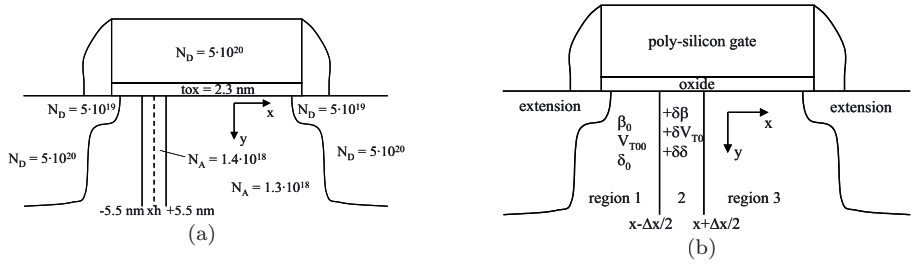


Figure 4.9. a) Schematical drawing of simulated MOSFETs with a slightly higher doping concentration between $x_h - 5.5$ nm and $x_h + 5.5$ nm. b) Schematical drawing of a MOSFET with a slightly higher threshold voltage, current factor or δ between $x - \Delta x/2$ and $x + \Delta x/2$.

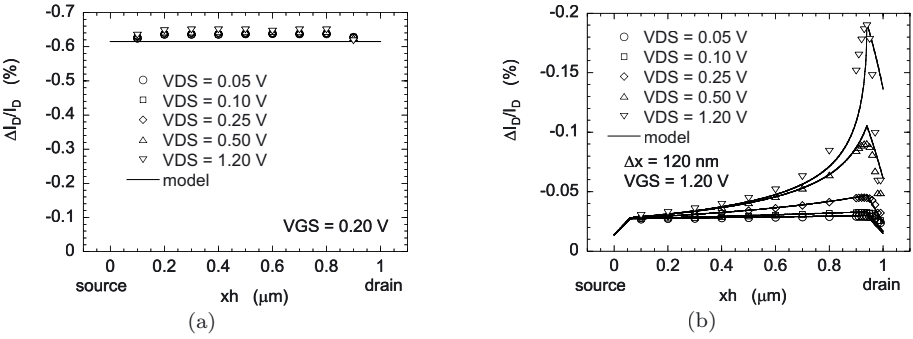


Figure 4.10. Simulated and calculated increase in the drain current as a function of the lateral position of a slight increase in the doping concentration as depicted in figure 4.9. a) Weak inversion. b) Strong inversion.

Now assume that $\delta\psi_s$ is normally distributed with mean 0 and variance $\sigma^2_{\delta\psi_s}$ and that its spacial distribution is described by a normalized power spectrum $f_{\delta\psi_s}(\omega_r)$. From this, it follows that δf_n has a lognormal distribution with mean $(\mu_{\delta f_n})$ and variance $(\sigma_{\delta f_n})$ equal to:

$$\mu_{\delta f_n} = e^{(q/kT)^2\sigma^2_{\delta\psi_s}/2} \tag{4.47}$$

$$\sigma^2_{\delta f_n} = e^{(q/kT)^2\sigma^2_{\delta\psi_s}}\left(e^{(q/kT)^2\sigma^2_{\delta\psi_s}} - 1\right). \tag{4.48}$$

Note that when $\sigma_{\delta\psi_s} \ll kT/q$, we can linearize the problem, i.e. $\delta f_n \cong 1 + (q/kT)\delta\psi_s$, and (4.47) simplifies to $\mu_{\delta f_n} \cong 1$ and (4.48) simplifies to $\sigma^2_{\delta f_n} \cong (q/kT)^2\sigma^2_{\delta\psi_s}$. However, in general this approximation will not be valid. The deviations from the ideal linear case will now be investigated.

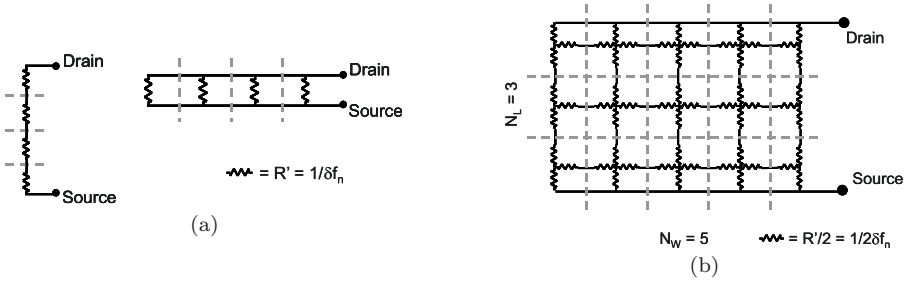


Figure 4.11. Representations of the MOSFET in weak inversion by resistor networks. a) Extreme situation of a very narrow or very short transistor. b) Both the length and width of the transistor are larger than the correlation length $l_{\delta\psi_s}$.

We will introduce the following nomenclature: The difference in a parameter P of a device that suffers from microscopic fluctuations and of an imaginary device without these fluctuations is denoted as $\Delta' P$. The mismatch between two macroscopically identical devices with only microscopic differences is denoted as ΔP. Furthermore, the correlation length of the mismatch causing stochastic process $f_{\delta P}(\omega_r)$ is defined as: $l_{\delta P} \equiv 2\pi\sqrt{f_{\delta P}(0)}$.
Consider the two extreme device shapes as depicted in figure 4.11a. In case of a very short device ($L \ll l_{\delta\psi_s}$) the macroscopic conductance is given by the sum of the local conductances, which for a wide enough transistor yields:

$$\mu_{\Delta' f_n} = \mu_{\delta f_n} - 1 \tag{4.49}$$

$$\sigma^2_{\Delta' f_n} = \frac{l_{\delta\psi_s}}{W}\sigma^2_{\delta f_n}. \tag{4.50}$$

When $\sigma_{\delta\psi_s}$ is a significant fraction of kT/q, the average current and the variation increase, which is mainly due to the exponentially higher conductance of local regions with high ψ_s.
Now consider the other extreme. For a very narrow device ($W \ll l_{\delta\psi_s}$) the macroscopic resistance is given by the sum of the local resistances, which for a long enough transistor yields:

$$\mu_{\Delta' f_n} = 1/\mu_{\delta f_n} - 1 \tag{4.51}$$

$$\sigma^2_{\Delta' f_n} = \frac{l_{\delta\psi_s}}{L}\sigma^2_{1/\delta f_n} = \frac{l_{\delta\psi_s}}{L}\frac{\sigma^2_{\delta f_n}}{\mu^4_{\delta f_n}}. \tag{4.52}$$

When $\sigma_{\delta\psi_s}$ is a significant fraction of kT/q, the average current and the variation decrease, which is mainly due to the exponentially higher re-

sistance of local regions with low ψ_s.
The relative current mismatch between two transistors $\Delta I_D/I_D = \Delta f_n/(1+\mu_{\Delta' f_n})$. From this it follows that $\Delta I_D/I_D$ has mean $\mu_{\Delta I_D/ID} = 0$ and variance:

$$\sigma^2_{\Delta I_D/I_D} = \frac{2\sigma^2_{\Delta' f_n}}{(1+\mu_{\Delta' f_n})^2} \tag{4.53}$$

According to the central limit theorem, $\Delta I_D/I_D$ has a normal distribution for transistors with large enough area.
Note that, until now, two dimensional current flows were neglected. A more realistic representation of the transistor is depicted in figure 4.11b. It consists of $N_W = W/l_{\delta\psi_s}$ segments in the width direction and $N_L = L/l_{\delta\psi_s}$ segments in the length direction. Each segment contains four resistors. In [3] a solution to this problem was found by considering each segment to be either switched on or off. A more accurate quasi-resistance approach was presented in [103]. We will follow a similar approach, but it is not attempted to analytically model the effect. To obtain the current variation, the resistor network out of figure 4.11b is simulated. The resistivity of each segment is calculated from a randomly assigned $\delta\psi_s$, that is taken from a normally distributed set. In order to reach high enough accuracy, 1000 microscopically different resistors were included.
Results of this exercise are presented in figure 4.12. Figure 4.12a shows the relative increase in current $(\mu_{\Delta' f_n} + 1)$ as a function of $\sigma_{\delta\psi_s}/(kT/q)$. Different line shapes represent different areas. The arrow indicates the direction of increasing N_W. Thus, for the $N_W N_L = 256$ case (full lines), the bottom line represents ($N_W = 2$, $N_L = 128$) and the top line represents ($N_W = 128$, $N_L = 2$). Figure 4.12b shows the same data as a function of N_W/N_L at $\sigma_{\delta\psi_s}/(kT/q) = 2$. Only for short transistors an increase is observed. For longer and wide transistors the current decreases with respect to the no-microscopic-fluctuations case. Figures 4.12c and 4.12d show the increase in $\sigma_{\Delta I_D/I_D}$ with respect to the results obtained with the linear approximation. At $\sigma_{\delta\psi_s}/(kT/q) = 2$ an increase of 30 % to 50 % is observed, which becomes larger for extreme N_W/N_L ratios. Note that, although one dimensional analysis yielded the same result for very short or narrow devices, it seriously overestimates the increase for the intermediate cases. This demonstrates that current tends to flow around regions of high resistivity, while it concentrates in regions of low resistivity. Finally, figures 4.12e and 4.12f plot the correlation between $\Delta' f_n$ and $\Delta' \psi_s$. This parameter is a measure of the correlation between current mismatch in weak and strong inversion. At $\sigma_{\delta\psi_s}/(kT/q) = 2$ its value is seen to be around 0.5 to 0.8, which is significantly smaller than 1.

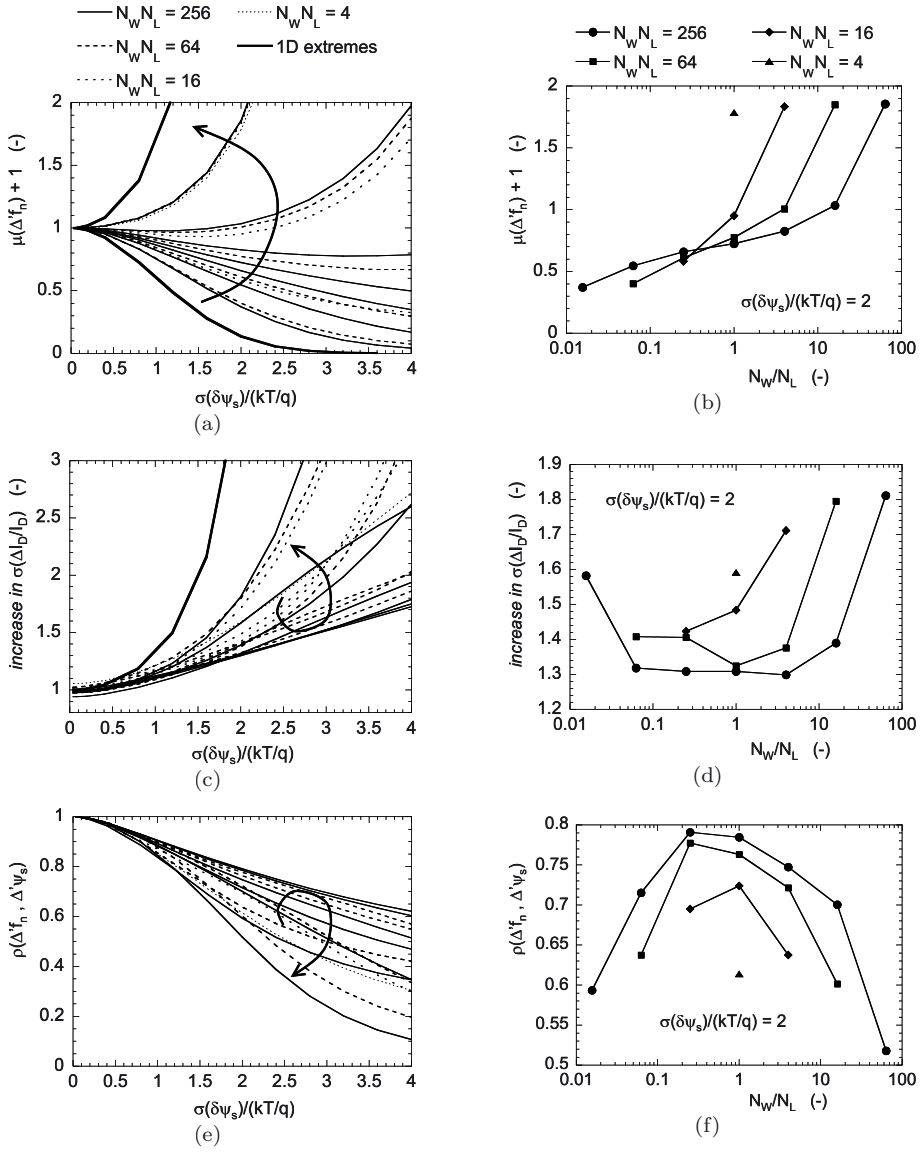


Figure 4.12. Simulation results of the resistor network out of figure 4.11b. Figures a+b show the increase in current, c+d the increase in standard deviation and e+f the correlation of current mismatch with $\Delta'\psi_s$. In figures a+c+e the results are plotted as a function of $\sigma_{\delta\psi_s}/(kT/q)$. The arrow indicates increasing N_W/N_L ratio. Figures b+d+f plot the results as a function of N_W/N_L for $\sigma_{\delta\psi_s}/(kT/q)$=2.

These simulation results can partly explain the differences between weak and strong inversion that were observed in the previous two chapters. Furthermore, it will be found in subsection 4.2.4 that short- and narrow-channel effects also give rise to significant differences.

4.2.2 Solution of the current equation in strong inversion

In strong inversion the mismatch in the drain current is due to mismatch in the current factor, mismatch in the threshold voltage and mismatch in δ as defined by (4.19). Their contributions will now be calculated. Consider a transistor that in between $x - \Delta x/2$ and $x + \Delta x/2$ has $\beta = \beta_0 + \delta\beta$, $V_{T0} = V_{T00} + \delta V_{T0}$ or $\delta = \delta_0 + \delta\delta$, while no other microscopic fluctuations are present (see figure 4.9b). This divides the transistor in three regions. In each region the potential V_{CS} is given by (4.21). However, the constant C is only equal to 0 in region 1. In the other regions it follows from the continuity of V_{CS}. Again, the drain current is found by using $V_{CS}(x = L) = V_{DS}$. To first order this results in[9]:

$$\left.\frac{\delta I_D}{I_D}\right|_{\delta\beta} = \frac{\Delta x}{L}\frac{\delta\beta}{\beta} \tag{4.54}$$

$$\left.\frac{\delta I_D}{I_D}\right|_{\delta V_{T0}} = \frac{-\beta}{I_D}\int_{x-\frac{\Delta x}{2}}^{x+\frac{\Delta x}{2}} \delta V_{T0}\frac{dV_{CS}}{dx'}dx' = \tag{4.55}$$

$$= \frac{-1}{L}\int_{x-\frac{\Delta x}{2}}^{x+\frac{\Delta x}{2}} \frac{\delta V_{T0}dx'}{V_{GS} - V_T - (1+\delta)V_{CS}(x')} \equiv \int_{x-\frac{\Delta x}{2}}^{x+\frac{\Delta x}{2}} \delta V_{T0}\cdot w'_{\delta V_{T0}}(x')dx'$$

$$\left.\frac{\delta I_D}{I_D}\right|_{\delta\delta} = \frac{-\beta(1+\delta)}{I_D}\int_{x-\frac{\Delta x}{2}}^{x+\frac{\Delta x}{2}} \delta\delta\cdot V_{CS}(x')\frac{dV_{CS}}{dx'}dx' = \tag{4.56}$$

$$= \frac{-1}{L}\int_{x-\frac{\Delta x}{2}}^{x+\frac{\Delta x}{2}} \frac{V_{CS}(x')\delta\delta dx'}{V_{GS} - V_T - (1+\delta)V_{CS}(x')} \equiv \int_{x-\frac{\Delta x}{2}}^{x+\frac{\Delta x}{2}} \delta\delta\cdot w'_{\delta\delta}(x')dx',$$

[9]For the mismatch in the threshold voltage, a similar analysis, but using somewhat different mathematics, was published in [41]. It was assumed that $\Delta x \to 0$. This approximation will turn out to be invalid. In parallel to our work, similar results were published in [25]. In this paper a logarithmic deviation to the $1/\sqrt{area}$ law was derived, which will turn out to be in accordance with our results.

where $w'_{\delta V_{T0}}(x)$ and $w'_{\delta\delta}(x)$ represent the sensitivity of $\Delta I_D/I_D$ to $\delta V_{T0}(x)$ and $\delta\delta(x)$, respectively. Outside the $0 < x < L$ interval these functions are put to 0.
It is seen that the impact of $\delta\beta$ is independent of its lateral position. Therefore, we can apply the analysis out of subsection 2.4.1 and it follows that:

$$\sigma^2_{\Delta I_D/I_D}|_{\Delta\beta} = \sigma^2_{\Delta\beta/\beta} = \frac{2l^2_{\delta\beta}\sigma^2_{\delta\beta/\beta}}{WL} \equiv \frac{A_{0,\Delta\beta/\beta}}{WL}, \tag{4.57}$$

where the correlation length $l_{\delta\beta} \equiv 2\pi\sqrt{f_{\delta\beta}(0)}$.
Now consider δV_{T0} and $\delta\delta$ and note that, from (4.14) and (4.19), they are expected to be fully correlated. Using the same kind of approximation as in (4.20) we can write:

$$\delta V_T = \left(1 + \delta_v \frac{V_{CS}}{\phi_B}\right) \delta V_{T0}, \tag{4.58}$$

where δ_v models the sensitivity of δV_T to V_{CS}/ϕ_B. The weighting functions $w_{\delta V_T}(x, \Delta x)$ and $w'_{\delta V_T}(x)$ are now defined as:

$$w_{\delta V_T}(x, \Delta x) = \int_{x-\frac{\Delta x}{2}}^{x+\frac{\Delta x}{2}} w'_{\delta V_T}(x)dx' = \tag{4.59}$$

$$= \int_{x-\frac{\Delta x}{2}}^{x+\frac{\Delta x}{2}} \left(w'_{\delta V_{T0}}(x') + \frac{\delta_v w'_{\delta\delta}(x')}{\phi_B}\right) dx'$$

Furthermore, it will turn out useful to define the following functions:

$$k_{V_T} = 1 + \frac{\delta_v}{1+\delta}\frac{V_{GS} - V_T}{\phi_B} \tag{4.60}$$

$$k_\delta = \frac{\delta_v}{(1+\delta)\phi_B}. \tag{4.61}$$

At high enough values of the drain bias, it follows from (4.55), (4.56) and (4.59) that a lateral dependence is expected of the sensitivity of the drain current to a local fluctuation in threshold voltage. It follows from (4.15) and (4.55) that this lateral dependence is approximately proportional to the lateral dependence of one over the inversion-layer charge or the local resistivity of the inversion layer.
To test (4.59), the simulations described by figure 4.9a are repeated

for strong inversion[10]. The mobility was taken as a constant in the simulations to avoid a change in current factor due to the change in doping. The results are plotted in figure 4.10b and are seen to be well described for $\Delta x = 120$ nm. It was assumed that $\delta_v = 0.5$. The value of δV_{T0} was calculated from the theory presented in section 4.1.1, but had to be multiplied by 1.1 to give a good fit. The magnitude of Δx seems too large to be caused by the introduced disturbance. A possible explanation is that the mismatch itself becomes smaller close to the source and drain regions. This could be caused by e.g. charge sharing. Taking this into consideration, the following weighting function will be assumed:

$$w'_{\delta V_T,fin}(x,\Delta x_s) = \begin{cases} 0 & x<0 \vee x>L \\ \frac{x}{\Delta x_s}\cdot w'_{\delta V_T}(\Delta x_s) & 0<x<\Delta x_s \\ w'_{\delta V_T}(x) & \Delta x_s < x < L-\Delta x_s \\ \frac{L-x}{\Delta x_s}\cdot w'_{\delta V_T}(L-\Delta x_s) & L-\Delta x_s < x < L \end{cases}, \tag{4.62}$$

where Δx_s is the range over which the extension regions affect δV_T. A similar fit as depicted in figure 4.10b is obtained when $\Delta x_s = 60$ nm. Generally, it can be assumed that $(V_{GS}-V_{T0})^2 \gg (L/\Delta x_s)(g_{out}/\beta)^2$, as follows from figure 4.3. The height of the peak is then fully determined by Δx_s, while the output conductance plays no significant role.
At low values of the drain bias the impact of a local disturbance of the threshold voltage on the drain current is given by:

$$\left.\frac{\delta I_D}{I_D}\right|_{\delta V_T} \cong \frac{-\delta V_{T0}}{V_{GS}-V_T}\frac{\Delta x}{L} \cong -\frac{g_m \delta V_{T0}}{I_D}\frac{\Delta x}{L}, \tag{4.63}$$

which is the same result as was obtained in subsection 2.3.2. This shows that, at low drain bias, all equations are linear and averaging effects can be interchanged.
The variance of the drain current is calculated as follows[11]:

$$\sigma^2_{\Delta I_D/I_D}|_{\Delta V_T} = \sigma^2_{\delta V_T}\left[\rho_{\delta V_T} * w'_{\delta V_T,fin} * w'_{\delta V_T,fin}\right](0)\frac{2l_{\delta V_T}}{W} \approx \tag{4.64}$$

$$\approx \sigma^2_{\delta V_T}\left[w'_{\delta V_T,fin} * w'_{\delta V_T,fin}\right](0)\frac{2l^2_{\delta V_T}}{W}$$

where the autocorrelation function $\rho_{\delta V_T}(x)$ describes the spacial properties of δV_T. It is equal to the Fourier transform of the normalized power

[10] A similar simulation was presented in [105], but only at low drain bias. In accordance with our result, no lateral dependence was observed.
[11] The symbol $*$ denotes the convolution integral: $[f_1 * f_2](x) \equiv \int_{-\infty}^{\infty} f_1(x')\cdot f_2(x-x')dx'$.

spectrum $f_{\delta V_T}(\omega_r)$. Note that (4.64) is identical to (2.38), but with an adapted geometry function that takes the weight of δV_T as a function of the lateral position into account. The last approximate equality in (4.64) is valid when the weighting function does not vary too rapidly over a distance of the correlation length $l_{\delta V_T}$, or, in other words, when $d^2 w'_{\delta V_T,fin}/dx^2 \ll w'_{\delta V_T,fin}(x)/l^2_{\delta V_T}$. This holds at low drain bias, or at high V_{DS} when $\Delta x_s \gg l_{\delta V_T}$ for all values of x.
At low drain bias (4.64) simplifies into:

$$\sigma^2_{\Delta I_D/I_D}|_{\Delta V_T} \cong \frac{1}{(V_{GS}-V_T)^2}\frac{2l^2_{\delta V_T}\sigma^2_{\delta V_T}}{WL} \cong \left(\frac{g_m}{I_D}\right)^2 \sigma^2_{\Delta V_T}, \qquad (4.65)$$

which, again, is the same result as was obtained in subsection 2.3.2. However, at higher drain bias, assuming $L \gg \Delta x_s \gg l_{\delta V_T}$:

$$\sigma^2_{\Delta I_D/I_D}|_{\Delta V_T} \approx \left(\frac{\ln\left(\frac{L}{\Delta x_s}\right)k^2_{V_T}}{(V_{GS}-V_{T0})^2} - \frac{4k_{V_T}k_\delta}{V_{GS}-V_{T0}} + k^2_\delta\right)\frac{2l^2_{\delta V_T}\sigma^2_{\delta V_T}}{WL}, \quad (4.66)$$

Applying the analysis out of chapter 2 instead of the analysis presented here, thus using (2.3) on (4.22) with $V_{DSsat} = (V_{GS}-V_{T0})/(1+\delta)$, yields:

$$\sigma^2_{\Delta I_D/I_D}|_{\Delta V_T} = \left(\frac{4k^2_{V_T}}{(V_{GS}-V_{T0})^2}\right)\frac{2l^2_{\delta V_T}\sigma^2_{\delta V_T}}{WL}, \qquad (4.67)$$

It is seen that (4.66) can give both smaller and larger results than (4.67)[12], depending on the length. Figure 4.13 plots the ratio of $\sigma^2_{\Delta V_T}(V_{DS} = V_{DD})$ and $\sigma^2_{\Delta V_T}(V_{DS} = 50\text{ mV})$ as a function of the gate length. The measurements were performed on the same technologies as described in subsection 4.1.2. In saturation, threshold-voltage mismatch was extracted from the mismatch in the drain current by applying (2.7) at $V_{GS} = V_T + 0.3$ V. Current-factor mismatch was neglected, which might cause small errors. The change in threshold voltage mismatch between the two regimes was calculated by dividing the right-hand side of (4.66) by the right-hand side of (4.67). Despite a lot of scatter on the experimental data, the increase of the ratio with length is significant. The long-channel transistors are seen to be reasonably well described by the model. Further justification will be presented in subsection 4.2.5,

[12]The difference between (4.66) and (4.67) has a similar origin as the difference between physical mobility and effective mobility, as used in most drain-current models. Generally, the drain current is derived by solving (4.2) and neglecting the bias dependence of the mobility. This bias dependence is only introduced in the solution for the drain current. A physically more correct approach introduces the bias dependence before solving (4.2).

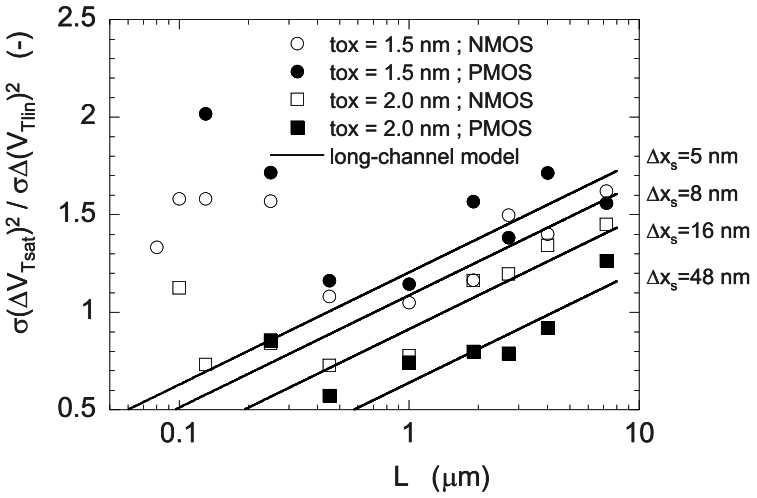


Figure 4.13. Ratio of $\sigma^2_{\Delta V_T}(V_{DS} = V_{DD})$ and $\sigma^2_{\Delta V_T}(V_{DS} = 50\text{ mV})$ as a function of the gate length.

where the symmetry of the MOSFET will be examined. The magnitude of Δx_s is found to be in the order of $5 - 50$ nm. The highest Δx_s is observed for the $t_{ox} = 2.0$ nm PMOS transistors, that also suffer most severely from short channel effects (see subsection 4.1.2). For the observed values of Δx_s, the approximation $\Delta x_s \gg l_{\delta V_T}$ is not expected to be fully valid. Therefore, besides short channel-effects, Δx_s is expected to be partly determined by the correlation length of the mismatch causing stochastic process.
The model is seen to be invalid for short devices. This is due to the fact that the equations for V_{CS}, inserted in (4.55) and (4.56), are incorrect. For short devices dV_{CS}/dx becomes independent of the lateral position due to velocity saturation. This implies that we can again safely use the equations presented in chapter 2. However, note that $\sigma_{\Delta V_T}$ itself is expected to vary with V_{DS} when short-channel effects become too severe. Again looking at (4.66) and (4.67), it is seen that at higher gate bias they do not converge to 0, which is due to the non-zero value of δ. This was neglected in subsection 2.3.2 and it explains the correlation between ΔV_T and $\Delta(1/\zeta_{sat})$ as observed in subsection 2.4.3, table 2.3 and figure 2.15a.

4.2.3 Short- and narrow-channel effects

As was already observed in section 2.4, deviations from the $\sigma^2_{\Delta P} \propto 1/WL$ law are expected for short or narrow transistors.

Short-channel effects. For short devices deviations can be due to 1) a smaller effective channel length than the metallurgical channel length [37, 38, 51], 2) the increase in surface potential, caused by the proximity of the extension regions [17, 31, 61], 3) the increase in doping level due to the halos [22, 75, 88, 92, 94][13] and 4) fluctuations in the short-channel effects themselves [95]. The second effect can cause a decrease in threshold-voltage mismatch, while the other effects increase the mismatch. In this subsection mainly the contribution of the lower effective channel length is investigated. The dependence of MOSFET mismatch on technology related parameters will be investigated in section 4.3 and chapter 5. To describe the impact of the smaller effective channel length, the following model is tried:

$$\sigma_{\Delta P} = \frac{A_{0,\Delta P}}{\sqrt{W(L - \Delta L_{\Delta P})}}, \tag{4.68}$$

where $\Delta L_{\Delta P}$ models the change in channel length. Figures 4.14a+b present fits to experimental data for threshold-voltage mismatch and current-factor mismatch, respectively. The values for $\Delta L_{\Delta P}$ are listed in table 4.1. The NMOS devices are seen to be well described by the model. Their short-channel effects are well controlled, as was earlier observed in subsection 4.1.2. The results for the PMOS transistors are less well fitted. The short devices suffer severely from short-channel effects and $\Delta L_{\Delta P}$ becomes a function of the gate length. When trying to describe short-channel effects, it is better to use (2.41) as opposed to (4.68), to avoid singularities. However, note that (4.68) has a more physical base. Figure 4.14c plots the ratio of $\sigma_{\Delta V_T}$ and $\sigma_{\Delta\beta/\beta}$. This ratio is seen to be constant for NMOS devices, which indicates that the main short-channel effect is the reduction in channel length. The difference in the extracted $\Delta L_{\Delta V_T}$ and $\Delta L_{\Delta\beta/\beta}$ is due to scatter on the experimental data. For the PMOS devices, also other effects are seen to play a role. Finally note that in figure 4.13 an increase of threshold voltage mismatch with increasing drain bias was observed for short transistors. This can be explained by the decrease in effective channel length, as described by (4.8), (4.25) and (4.29) and as reported in literature [51, 81].

Narrow-channel effects. To describe the impact of narrow-channel effects on $\sigma_{\Delta P}$, we will make the same approximation as in subsection 4.1.2: The device is assumed to consist of three transistors in parallel, namely one center transistor and two transistors at the side. The transistors at the side can have different threshold voltage and also $\sigma_{\Delta V_T}$

[13] or other structural changes related to short-channel devices

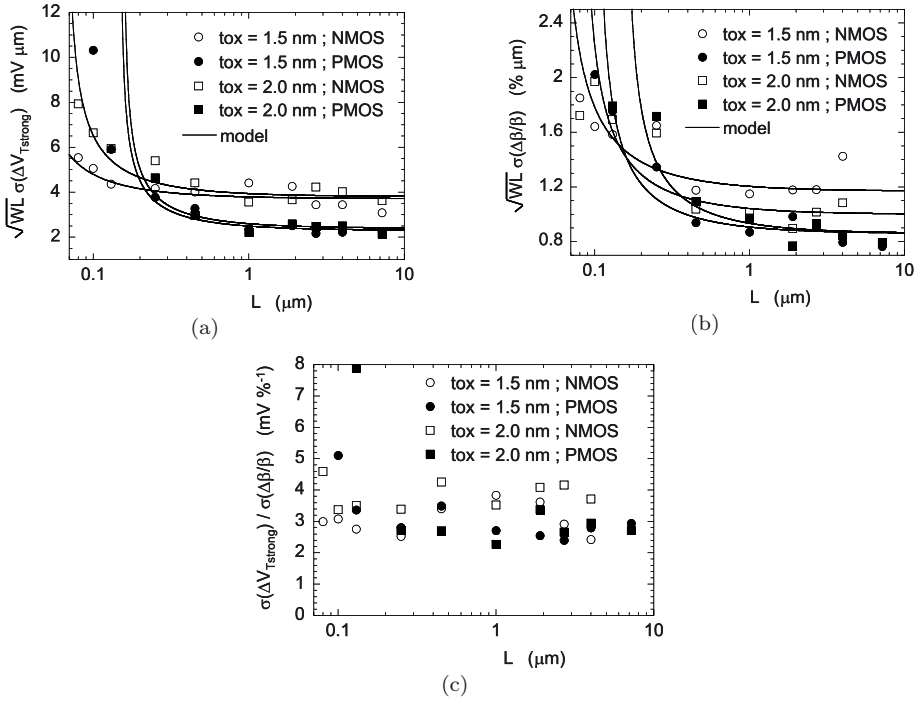


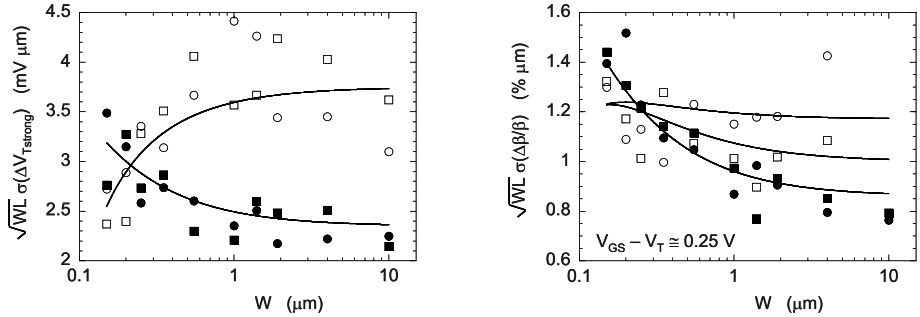
Figure 4.14. $\sqrt{WL}\sigma_{\Delta V_T}$ (a), $\sqrt{WL}\sigma_{\Delta\beta/\beta}$ (b) and their ratio (c) as a function of the transistor length. $V_{DS} = 50$ mV. Symbols have the same meaning as in figure 4.13.

and $\sigma_{\Delta\beta/\beta}$ are expected to differ. The overall variation is determined by averaging the fluctuations over the width of the transistor. The increase in current density at the edges needs to be taken into account. In general, this gives:

$$\sigma^2_{\Delta P} = \frac{\left(\frac{dI_{Dmiddle}}{dP}\right)^2 \frac{A^2_{0,\Delta P}}{W_{middle}L} + 2\left(\frac{dI_{Dnarrow}}{dP}\right)^2 \frac{A^2_{narrow,\Delta P}}{W_{narrow}L}}{\left(\frac{dI_{Dmiddle}}{dP} + 2\frac{dI_{Dnarrow}}{dP}\right)^2}, \qquad (4.69)$$

where $I_{Dmiddle}$ and $I_{Dnarrow}$ are the current flowing in the middle and edge transistors, respectively. The variation at the edge is described by $A^2_{narrow,\Delta P}/W_{narrow}L$. Consider the linear regime as example. For this regime (4.69) yields:

$$\sigma^2_{\Delta V_T} = \frac{W_{middle}}{W}\frac{A^2_{0,\Delta V_T}}{WL} + \frac{2W_{narrow}}{W}\frac{A^2_{narrow,\Delta V_T}}{WL}, \qquad (4.70)$$



(a) (b)

Figure 4.15. $\sqrt{WL}\sigma_{\Delta V_T}$ (a) and $\sqrt{WL}\sigma_{\Delta\beta/\beta}$ (b) as a function of the transistor width. $V_{DS} = 50$ mV. Symbols have the same meaning as in figure 4.14.

$$\sigma^2_{\Delta\beta/\beta} = \frac{W_{middle}}{W}\left(\frac{V_{GS}-V_{Tlw}}{V_{GS}-V_T(W)}\right)^2 \frac{A^2_{0,\Delta\beta/\beta}}{WL} + \tag{4.71}$$

$$+ \frac{2W_{narrow}}{W}\left(\frac{V_{GS}-V_{Tnarrow}}{V_{GS}-V_T(W)}\right)^2 \frac{A^2_{narrow,\Delta\beta/\beta}}{WL}.$$

Figure 4.15 shows that these models give a good description of experimental data. Model parameters are listed in table 4.1. For narrow NMOS transistors $\sqrt{WL}\cdot\sigma_{\Delta V_T}$ is seen to be smaller than for wide transistors. This could be explained by a lower doping level at the edge of the transistor, since this would result in reduced doping fluctuations. For the PMOS devices the doping level is more or less constant and STI is seen to increase $\sqrt{WL}\cdot\sigma_{\Delta V_T}$ for more narrow transistors. Also an increase in $\sigma_{\Delta\beta/\beta}$ is observed for the PMOSFETs, which could be caused by sidewall roughness. This effect is seen to be less prominent for the NMOSFETs.

4.2.4 Comparison of mismatch in weak and strong inversion

The analysis presented in the previous subsection is valid in strong inversion. Deviations in weak inversion will now be discussed. Firstly consider the effect of halos. It was found in subsection 4.2.2 that the weight attributed to the local fluctuations is inversely proportional to local value of the inversion layer charge. From this it follows that:

$$\sigma^2_{\Delta V_T} = \frac{L\int_0^L \frac{dx}{Q_i(\psi_s(x))^2}}{\left(\int_0^L \frac{dx}{Q_i(\psi_s(x))}\right)^2}\frac{A^2_{0,\Delta V_T}}{WL}. \tag{4.72}$$

This equation is valid in weak inversion. In strong inversion $\psi_s(x)$ needs to be replaced by:

$$V_T(x) = V_{Tlw} - (1+\delta)(\psi_s(x) - \psi_s^0). \tag{4.73}$$

The surface potential is calculated with (4.24). Note that we have assumed that the fluctuation mechanism doesn't change with lateral position x. In strong inversion, a not too strong halo mainly reduces the short-channel effect. The variation of ψ_s with x is not too large and can be neglected[14]. However, in weak inversion halos are expected too play a more significant role, because of the exponential dependence of inversion-layer charge to surface potential.

Figure 4.16a compares the calculation of the ratio of $\sigma^2_{\Delta V_T}$ in weak[15] and strong inversion to experimental data. The parameters out of table 4.1 were used in the calculation. For long-channel NMOS devices, it is observed that an increase in the ratio is expected. This can be explained from figure 4.5b. The halos cause two bumps in the surface potential profile. For short transistors these bumps overlap and the only difference between weak and strong inversion is in effective channel length. For increasing length, the bumps appear and gain in relative importance. For very long transistors the impact of the halos is expected to decrease again. For the PMOS transistors the halos were found to be less effective (see figure 4.6a) and it follows from the calculation that no significant increase is expected. In figure 4.16a, it is observed that (4.72) underestimates the experimental data. This could be partly related to an inaccurate estimate of the surface-potential profile. However, also the width dependence has been neglected. It will turn out to explain most of the experimentally observed differences. For the short $t_{ox} = 1.5$ nm PMOS transistors, also an increase of the weak inversion mismatch is observed. This could be due the decrease in L_{eff}, but it is not observed for the other technologies.

Equation (4.69) will be used to investigate the impact of the narrow-channel effect on the difference between weak and strong inversion. However, first note that the width of the edge transistor can be different in the two regimes, as is illustrated in figure 4.17. When the change in threshold voltage in width direction (z) is abrupt, $W^{weak}_{narrow} = W^{strong}_{narrow}$. For a non-abrupt change, $W^{weak}_{narrow} < W^{strong}_{narrow}$, due to the exponential dependence of drain current on surface potential. In case of a trapezoidal profile $W^{weak}_{narrow} \cong (1+\delta)(kT/q)W^{strong}_{narrow}/(V_{Tlw} - V_{Tnarrow})$.

[14]Experimentally the impact of halos on device mismatch will be more thoroughly investigated in section 5.3.

[15]In weak inversion, threshold voltage mismatch is equal to the mismatch in gate bias.

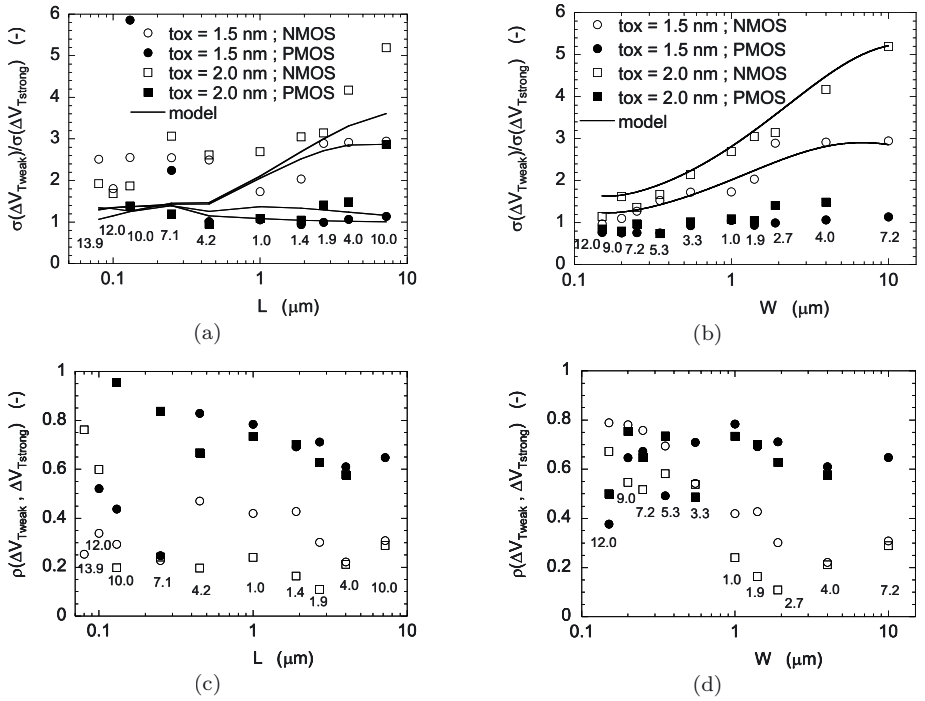


Figure 4.16. a+b) Ratio of $\sigma^2_{\Delta V_T}$ in weak inversion and $\sigma^2_{\Delta V_T}$ in strong inversion as a function of the transistor length (a) and width (b). c+d) Correlation between ΔV_T in weak and strong inversion as a function of the transistor length (c) and width (d). Device widths (a+c) or lengths (b+d) are included in the figures and are given in μm. $V_{DS} = 50$ mV.

Figure 4.16b shows the calculated and experimental increase of $\sigma^2_{\Delta V_T}$ as a function of the width. Model parameters are again taken from table 4.1. Their values were obtained by a combined fit to the experimental results presented in figures 4.6b, 4.15a and 4.16b. Good agreements between the fits and experimental data are observed. However, note that in this case the impact of the halos was neglected.

To explain these results, the same kind of reasoning follows as earlier. For narrow transistors, the device consists mainly of the edge transistor and no difference is expected. For wider transistors, the weak inversion $\sigma^2_{\Delta V_T}$ is still mainly determined by the edge transistors, due to the enormously larger edge-current density. In other words, the effective width of the device is reduced, which causes an increase in the variation. In strong inversion, this effect is much less pronounced. As expected, the PMOS transistors do not suffer from this increase in threshold voltage, since they don't possess a lower threshold voltage at the edge.

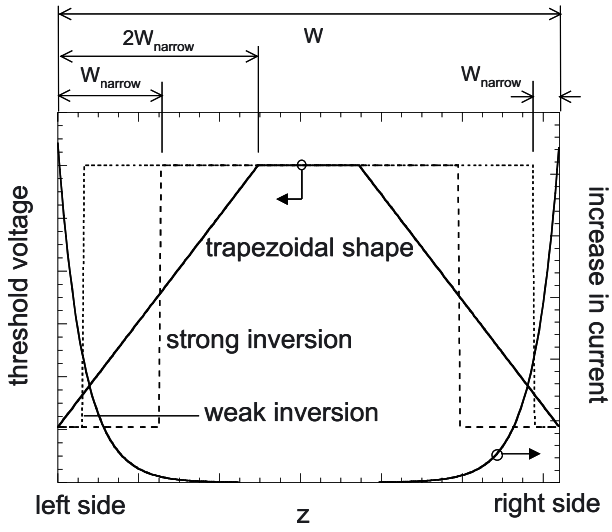


Figure 4.17. Schematic drawing of the threshold voltage (left axis) as a function of its position in the width direction (z). On the right axis the increase in drain current for the trapezoidal V_T profile is plotted.

Finally consider figures 4.16c+d, that show the correlation between ΔV_T in weak and strong inversion as a function of the gate length and width, respectively. As expected, the correlation is seen to decrease with increasing width. It seems to be less sensitive to the length, which indicates that the edge effect is more significant for the technologies under test. The correlation factor has a maximum value of ~ 80 %, which is significantly lower than 100 %. This could be due to the percolation effects that were observed in subsection 4.2.1, but that have been neglected in this subsection. The halo creates a sharp potential peak at the source and drain sides. Local increases in potential can cause current paths through this barrier, that decrease its impact[16]. In the edge transistor the current can be blocked by local regions of low potential. Again, this will decrease the impact. In order to describe the mixture of all these effects, detailed knowledge of the mismatch causing stochastic processes is required. Next, the problem has to be translated into a resistor network problem, taking into account the influence of halos and edge effects. Another option is to make use of a 3D simulator. However, this approach could be very time consuming.

[16]In [94], this effect was observed by comparing 2D and 3D simulations.

4.2.5 Asymmetry of MOSFET mismatch

To end this section, we will investigate the asymmetry of MOSFET mismatch. This creates extra insight in the position dependence of the impact of microscopic fluctuations on macroscopic parameters [13, 106]. In addition to literature we will demonstrate that asymmetry is also present for long transistors, and that it increases with increasing length. This is directly linked to the deviation from the $1/\sqrt{area}$ law, as was observed in subsection 4.2.2.
The asymmetry in the mismatch of the drain current is defined as:

$$asymmetry = \frac{\sigma^2(I_{D2f} - I_{D1f} - I_{D2r} + I_{D1r})}{\sigma^2(I_{D2f} - I_{D1f}) + \sigma^2(I_{D2r} - I_{D1r})}, \tag{4.74}$$

where the subscript 1 or 2 denotes the transistor number, the subscript f means that the transistor is measured with normal source and drain definitions, while the subscript r means reversed source and drain definitions. From the analysis in subsection 4.2.1 no asymmetry is expected for long transistors in weak inversion. For short transistors at higher drain bias, asymmetry in the lateral surface-potential profile (see figure 4.4b) can cause asymmetry in the current. For long-channel transistors in strong inversion, at high enough values of the drain bias, asymmetry in the drain current is expected due to asymmetry in the inversion-layer charge-density. From (4.74) and the analysis in subsection 4.2.2 it is expected to be equal to:

$$asymmetry(\Delta x_s) = \frac{\int_0^L (w'_{\delta V_T,fin}(x, \Delta x_s) - w'_{\delta V_T,fin}(L - x, \Delta x_s))^2 dx}{2L \cdot \sigma^2_{\Delta I_D/I_D}}. \tag{4.75}$$

For short transistors the inversion layer is expected to become less asymmetrical. However, channel-length modulation could cause extra asymmetry.
Figure 4.18 shows experimentally obtained curves of the asymmetry as a function of the gate bias for four different values of the drain bias. The presented results are for the $t_{ox} = 1.5$ nm NMOS transistors, but similar results were obtained for the other cases. In order to determine the asymmetry, we had to measure each transistor separately, instead of using the measurement algorithm described in subsection 2.1.3. Measurement repeatability was checked, and found to be no issue. However, as a side effect of using a different measurement algorithm, the source current was measured instead of the drain current.
At a drain bias of $V_{DS} = 100$ mV (figure 4.18b) no significant asymmetry is observed in as well weak as strong inversion. In moderate inversion

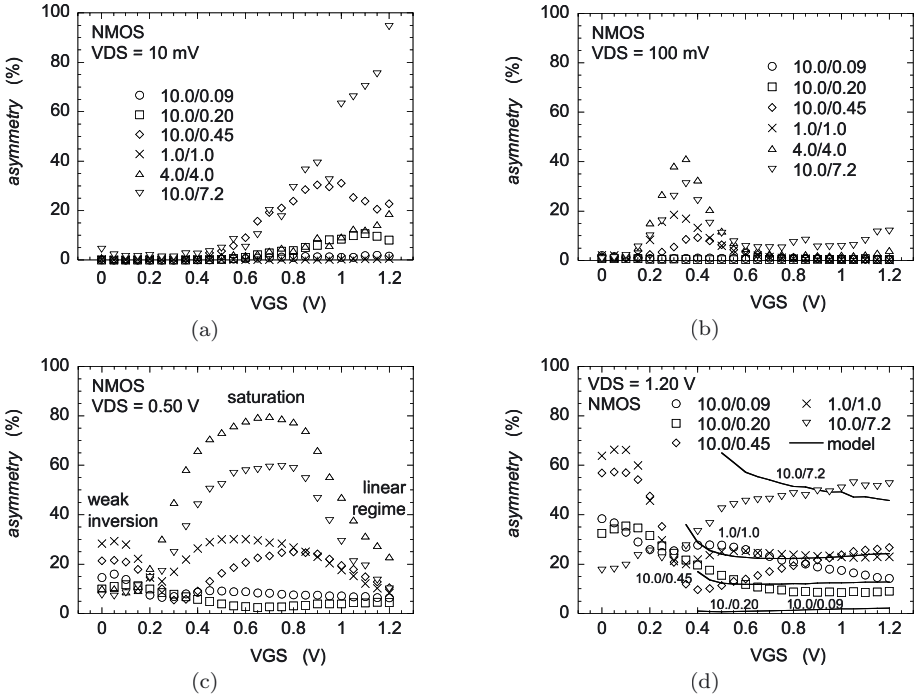


Figure 4.18. Asymmetry for four different values of the drain bias. The gate length is used as a parameter. The width over length ratio is given in μm/μm.

the device becomes asymmetric. In this regime the transistor operates in weak inversion at the drain side, while at the source side it is in saturation. This is a highly asymmetric situation.

At very low drain bias ($V_{DS} = 10$ mV) and high gate bias, asymmetry was observed for some of the examined pair dimensions. This could be attributed to a measurement issue. One transistor in a module shares its source[17] with a lot of other transistors. Therefore, at high gate bias, the total amount of tunnelling current through the source-gate overlap capacitance becomes significant as compared to the low drain current at low drain bias. Drains are connected separately. Therefore, the reverse measurement does not suffer from this problem.

Now consider the intermediate drain bias ($V_{DS} = 0.50$ V) case. At high gate bias, the transistors operate in the linear regime and are seen to be symmetric. Lowering the gate bias, moves the transistors into saturation, and causes asymmetry, which is most prominent for long channel

[17] in the case of normal terminal definitions

transistors. The high asymmetry for the $W = L = 4.0$ μm devices is not fully understood, but might be caused by inaccuracies due to limited statistics[18]. In weak inversion the transistors are again symmetrical, although signs of asymmetry start to be observed.
Finally consider the high drain bias case ($V_{DS} = 1.20$ V). In strong inversion, the transistors are fully operating in saturation. For long transistors, the model is seen to give a reasonable description of the experimental data for $\Delta x_s = 75$ nm. This value is higher than the value found in subsection 4.2.1, which was $\Delta x_s = 8$ nm. Note that both the measurements out of figure 4.13 and out of figure 4.18 are not very accurate and that we neglected current-factor mismatch. Furthermore, in the next section it will be found that most mismatch causing mechanisms contain both a threshold-voltage and current-factor component. This is not properly accounted for. However, the predictions of our model are qualitatively verified.
Finally consider the short transistors, for which asymmetry due to channel length modulation is observed. In weak inversion, the asymmetry is most prominent for intermediate gate lengths. This can only be explained by the impact of the halos. The drain bias lowers the barrier due to the halo at the drain side, while the halo barrier at the source side remains fully intact. Similar behavior was experimentally observed and simulated in [107].

4.3 Physical origins of fluctuations

This section describes the fluctuation mechanisms that cause the variability of MOS transistors and it calculates their impact. The main origins of the fluctuations are identified as: 1) doping fluctuations in the channel, 2) doping fluctuations in the gate, 3) fluctuations in the oxide charge, and 4) fluctuations in surface-roughness scattering. The magnitude of the doping fluctuations and oxide charge will be modeled in subsection 4.3.1 following the work published in [3]. Our calculation of the magnitude of the fluctuations in surface-roughness scattering is presented in subsection 4.3.5 and it is based on the statistical properties of the oxide-silicon interface.
The four fluctuation mechanisms affect transistor operation by influencing: 1) the threshold voltage, 2) the amount of gate depletion, 3) the magnitude of quantum-mechanical effects, and 4) the mobility. The impact of the fluctuation mechanisms on these macroscopic transistor

[18] Due to a measurement problem, only 42 device pairs per geometry were measured successfully. The other experimental results presented in this chapter are based on 84 measured device pairs per geometry.

parameters is calculated by using the charge-sheet approximation, i.e. it is again assumed that:

$$\Delta' N_{dope}(y)dy = \frac{dy}{WL}\int_{area} \delta N_{dope}(x,y,z)dxdz, \tag{4.76}$$

where $N_{dope}dy$ is related to the channel or gate doping or any other charge sheet in the transistor[19]. This equation was shown to be valid in strong inversion at low drain bias. Deviations in other regimes and for short- and narrow-channel transistors were discussed in the previous section. In order to calculate the variation in macroscopic parameter P, the fluctuations in the charge sheets need to be averaged out over the depth of the transistor:

$$\sigma^2_{\Delta P} = \int_{depth} \left(\frac{dP}{dN_{dope}(y)}\right)^2 \sigma^2_{\Delta N_{dope}}(y)dy. \tag{4.77}$$

The first factor in the integral models the sensitivity of P to a fluctuation in the doping at depth y. This sensitivity can be determined by e.g. simulations or modeling. Here, the modeling approach is followed, since it creates insight in the origin of the sensitivity. Higher accuracy might be obtained by using simulations.
We will start our calculations by examining one of the best studied mismatch phenomena, namely the impact of doping fluctuations on threshold voltage mismatch (subsection 4.3.2). The following subsections deal with gate depletion (subsection 4.3.3), quantum-mechanical effects (subsection 4.3.4), and mobility fluctuations (subsection 4.3.5). Subsection 4.3.6 combines all these effects in one model in order to determine which of them are relevant and to make a comparison to experimental data. The physical content of our models will be tested by examining gate- and bulk bias dependencies. Finally, in subsection 4.3.7 the results will be discussed. We will find that our calculations provide results that are close to the experimentally observed mismatch, while only two unknown parameters, related to gate depletion and surface-roughness scattering, need to be fitted. Finally, note that the presented equations related to quantum-mechanical effects and mobility are valid for NMOS transistors. For PMOS transistors appropriate changes in proportionality constants are required.

4.3.1 Doping fluctuations

Variation in the amount of dopants can be caused by numerous effects. Overall, the probability (p) that one dopant is present in a small

[19]For surface-roughness scattering a similar equation applies.

volume element (dV) is equal to $p = N_A dV$, independent of the presence of dopants in other volume elements. When the volume element is taken small enough, the chance that two dopants are present is negligible. The total number of dopants (N_{Atot}) in a certain volume (V) is then Poisson distributed[20] with mean N_{Atot} and variance $\sigma^2_{N_{Atot}} = N_{Atot}$. It follows that the doping concentration N_A is also Poisson distributed with mean N_A and variance $\sigma^2_{N_A} = N_A/V$. In a MOSFET, the average distance between dopants is of the order of magnitude of ~ 10 nm. This means that in most practical cases the Poisson distribution can be approximated by a normal distribution with the same mean and variance.
The same kind of analysis holds for dopants in the gate. However, note that the variance can be higher due to the poly-grain structure of the gate material. This will be further looked into in subsection 4.3.3 and section 5.2.
Finally note that besides the implanted channel doping, extra charge sheets (Q_{cs}) can exist, due to e.g. interface states, oxide charge or boron penetration. Using the same statistics, one can write $\sigma^2_{Q_{cs}} = q|Q_{cs}|/WL$. In case of boron penetration the variance is again expected to be more related to the randomness of the gate structure than to number fluctuations and it will be higher.

4.3.2 Impact of fluctuations in channel doping on threshold voltage

Doping fluctuations are considered to determine the lower obtainable limit to the variation of MOSFET parameters. Therefore, random dopant effects have been extensively studied in literature [10–12, 14, 51, 75, 77–81, 83, 84, 86–95, 105, 108]. To describe the impact of doping fluctuations, the same approach as in [10, 11, 76] will be followed. Furthermore, we will estimate the correlation length related to threshold voltage fluctuations due to random dopants. Finally the impact of the doping profile is examined.
We will now calculate the impact of doping fluctuations on the gate potential for fixed surface potential. In strong inversion this equals the mismatch in threshold voltage. Consider a charge sheet with thickness dy at a distance y from the interface that has a doping concentration that is $\delta N_A(y)$ higher than the average N_A. It follows from (4.6) that this increase results in such a decrease in depletion-layer width, that at its edge an amount of charge equal to $(y/W_D)\delta N_A(y)dy$ is covered. In

[20] In [14] it was shown that clustering of dopants increases the variance by the average amount of dopants that are clustered together. We will assume that no clustering takes place.

total, the shift in gate bias (δV_{GS}) due to $\delta N_A dy$ equals:

$$\delta V_{GS} = \frac{q t_{ox}}{\epsilon_{ox}} \left(1 - \frac{y}{W_D}\right) \delta N_A dy. \tag{4.78}$$

From this the variance of $\Delta' V_T$ ($\sigma^2_{\Delta' V_T}$) follows directly:

$$\sigma^2_{\Delta' V_T} = \frac{q^2 t^2_{ox}}{W L \epsilon^2_{ox}} \int_0^{W_D} \left(1 - \frac{y}{W_D}\right)^2 N_A(y) dy = \frac{t^2_{ox} \sqrt{2 q^3 \epsilon_{si} N_A \psi_s}}{3 W L \epsilon^2_{ox}}. \tag{4.79}$$

The last equality holds for uniform doping profiles. Note that this equation was derived considering only one device and that $\sigma^2_{\Delta V_T} = 2\sigma^2_{\Delta' V_{GS}}$. Besides the implanted channel doping, extra charge sheets (Q_{cs}) can exists, due to e.g. interface states, oxide charge or boron penetration. From (4.78), it follows directly that:

$$\sigma^2_{\Delta' V_T} = \frac{q |Q_{sc}| t^2_{ox}}{W L \epsilon^2_{ox}} \left(1 - \frac{y_{cs}}{W_D}\right)^2, \tag{4.80}$$

where y_{cs} is the depth of the charge sheet. For modern-day heavily-nitrided gate oxides, the fixed oxide-charge density can be as high as $N_f = 2 \cdot 10^{11}$ cm^{-2}. Although not dominant, its contribution cannot be neglected[21]. Note that this situation might worsen once high-k dielectrics are introduced.

Using (4.79), figure 4.19 shows the calculated threshold-voltage fluctuations as a function of the effective oxide thickness for transistors with minimum dimensions ($WL = 3L^2_{gate}$) and for transistors with $W = L = 1.0$ μm. Technology parameters are taken out of the ITRS roadmap [109], and listed in table 4.2. For devices with a constant area, the variation lessens for each technology generation due to the decrease in oxide thickness. Clearly, this is advantageous for analog operation. However, it is also seen that the variation of the minimum device increases when technologies are scaled down[22]. Since some modern-day digital circuits can contain an enormous amount of transistors ($> 10^9$), the requirements on $\sigma_{\Delta V_T}$ are quite stringent. From figure 4.19 it is clear that parameter variations are starting to play an important role in determining the design rules for digital circuits (see e.g. [7, 8]). Note again that doping variations give the lower limit to parameter fluctuations. In

[21] The simulations presented in [85] showed that the effects of fluctuations in the fixed oxide-charge can be neglected. However they considered $N_f = 2 \cdot 10^{10}$ cm^{-3}, which is low for heavily-nitrided gate oxides.

[22] In reality the variations are even slightly higher since $L_{eff} < L_{gate}$.

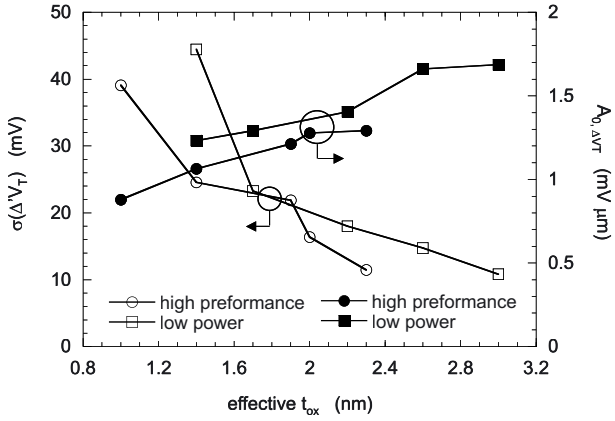


Figure 4.19. Threshold-voltage fluctuations as a function of the effective oxide thickness for transistors with minimum dimensions ($WL = 3L_{gate}^2$) (open symbols, left axis) and for transistors with $W = L = 1.0$ μm (filled symbols, right axis). Technology parameters are listed in table 4.2

Table 4.2. Technology parameters out of the ITRS roadmap that are used in the calculations presented in figure 4.19

high performance					
$t_{ox,eff}$ (nm)	2.3	2.0	1.9	1.4	1.0
L_{gate} (nm)	65	45	32	25	13
N_A (cm^{-3})	$1.5 \cdot 10^{18}$	$2.5 \cdot 10^{18}$	$2.5 \cdot 10^{18}$	$5.0 \cdot 10^{18}$	$9.0 \cdot 10^{18}$
low power					
$t_{ox,eff}$ (nm)	3.0	2.6	2.2	1.7	1.4
L_{gate} (nm)	90	65	45	32	16
N_A (cm^{-3})	$1.5 \cdot 10^{18}$	$2.5 \cdot 10^{18}$	$2.5 \cdot 10^{18}$	$5.0 \cdot 10^{18}$	$9.0 \cdot 10^{18}$

practice the fluctuations will be larger.
We will proceed this subsection by estimating the correlation length $l_{\delta\psi_s}$ and variance $\sigma^2_{\delta\psi_s}$, as defined in subsection 4.2.1. The impact on the surface potential ($V_{q_p}(y,r)$) of a point charge (q_p) at a distance y from the interface is approximated by:

$$V_{q_p}(y,r) = \frac{q_p}{4\pi\epsilon_{si}} \left(\frac{1}{\sqrt{y^2+r^2}} - \frac{1}{\sqrt{(y+2(\epsilon_{si}/\epsilon_{ox})t_{ox})^2+r^2}} \right). \quad (4.81)$$

The first term inside the brackets is due to the charge itself, the second term is due to its mirror charge in the gate. Assume that a dopant at depth y influences the surface potential over an area $l'^2_{\delta\psi_s}(y)$ of:

$$l'^2_{\delta\psi_s}(y) = \frac{1}{V_{q_p}(y,0)} \int_0^\infty 2\pi r V_{q_p}(y,r)dr = 2\pi y(y + 2(\epsilon_{si}/\epsilon_{ox})t_{ox}). \quad (4.82)$$

As rough estimate for $l_{\delta\psi_s}$ one can now use:

$$l^2_{\delta\psi_s} \sim \frac{\int_0^{W_D}(1 - \frac{y}{W_D})^2 N_A(y) \cdot l'^2_{\delta\psi_s}(y)dy}{\int_0^{W_D}(1 - \frac{y}{W_D})^2 N_A(y) \cdot dy} = \frac{\pi}{5}W_D^2 + \frac{\pi\epsilon_{si}}{\epsilon_{ox}}W_D t_{ox}. \quad (4.83)$$

The last equality is valid for a uniform doping profile. The variance $\sigma_{\delta\psi_s}$ is estimated by using (4.79) and putting $W = L = l_{\delta\psi_s}$. When $N_A = 1 \cdot 10^{18}$ cm^{-3} and $t_{ox} = 2.0$ nm, this gives $l_{\delta\psi_s} \sim 36$ nm and $\sigma_{\delta\psi_s} \sim 28$ mV $\approx kT/q$. In other words, the magnitude of the local variation in the surface potential is comparable to the thermal voltage. This means that to accurately predict the mismatch in weak inversion, 3D analysis is required, as was derived in subsection 4.2.1. For a device to fully operate in strong inversion, the gate overdrive needs to be significantly larger than $\sigma_{\delta\psi_s}$ ($V_{GS} - V_T \gtrsim 3\sigma_{\delta\psi_s}$). This is the case at the commonly used bias condition of $V_{GS} - V_T = 150$ mV.

In reality, the doping profile is not uniform. Figure 4.20 shows SIMS profiles of the doping concentration in the NMOS and PMOS transistors of a 0.13 μm technology with $t_{ox} = 2.0$ nm[23], $L_{nominal} = 130$ nm and $|V_{DD}| = 1.5$ V. These profiles are seen to be well described by Gaussian peaks:

$$N_A(y) = N_{A0} \cdot e^{-\left(\frac{y - D_{N_A}}{W_{N_A}}\right)^2}, \quad (4.84)$$

where N_{A0} is the peak concentration, D_{N_A} the peak position and W_{N_A} the width of the peak. Knowing the profile, W_D, $A_{0,\Delta V_T}$, $l_{\delta V_T}$ and $\sigma_{\delta V_T}$ can be calculated using (4.7), (4.79) and (4.83). Extracted and calculated values for all parameters are listed in table 4.3. Also listed are the values obtained by assuming a uniform doping profile. The uniform doping concentration is chosen in such a way, that the depletion layer charge at threshold equals that of the non-uniform case. Depending on the sharpness of the peak, these values are about 20 % to 35 % smaller

[23]This value is related to the physical oxide thickness. The effective (or electrical) oxide thickness is equal to $t_{oxeff} \cong 2.7$ nm. This value takes into account gate depletion and quantummechanical effects. These phenomena will be studied in a more decent way in subsections 4.3.3 and 4.3.4, respectively.

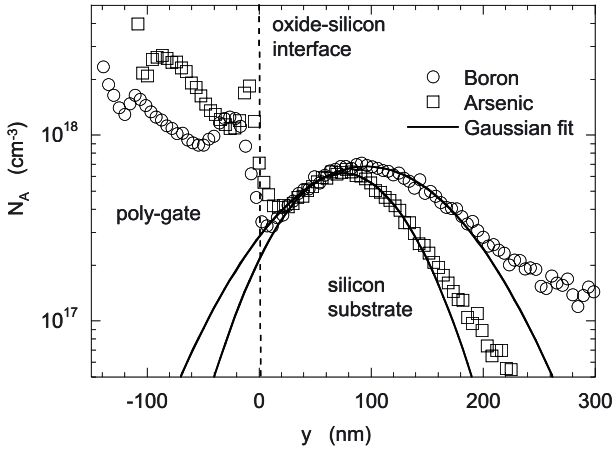


Figure 4.20. SIMS profiles of the boron and arsenic concentration for a 0.13 μm CMOS technology. The high concentrations of dopant atoms in the gate are caused by the halo implantations.

Table 4.3. Parameters related to doping fluctuations, calculated from the arsenic and boron SIMS profiles out of figure 4.20. These parameters are compared to calculations, that assume uniform doping profiles.

	Gaussian profile		uniform profile	
	boron	arsenic	boron	arsenic
N_A (cm^{-3})	$6.78 \cdot 10^{17}$	$6.25 \cdot 10^{17}$	$4.07 \cdot 10^{17}$	$3.63 \cdot 10^{17}$
D_{N_A} (nm)	96	75	-	-
W_{N_A} (nm)	103	72	-	-
W_D (nm)	52.3	53.3	56.5	59.8
$A_{0,\Delta V_T}$ (mVμm)	1.39	1.30	1.55	1.51
$l_{\delta V_T}$ (nm)	59.7	63.0	58.7	61.5
$\sigma_{\delta V_T}$ (mV)	16.5	14.6	18.7	17.3

than those obtained from the body coefficient, that is extracted by varying the bulk bias. In the table, it is observed that the differences are not very large between using the correct doping profiles or assuming a uniform doping concentration. The values for $A_{0,\Delta V_T}$ and $\sigma_{\delta V_T}$ are slightly lower and $l_{\delta V_T}$ is somewhat larger. However, using extreme retrograde doping profiles could significantly lower the fluctuations, since most of the doping is then moved away from the oxide-silicon interface. This

improvement follows directly from (4.79) and has been demonstrated by experiment [10] and simulations [75, 77, 80, 81, 84, 108].

4.3.3 Gate depletion

This subsection investigates the impact of gate depletion on MOSFET parameter fluctuations. It was shown in [21, 110–113] that gate depletion can severely degrade the matching performance. In [110–113] models are presented to describe the effect. However, the contribution of the inversion layer charge was not taken into account. The simulations of [110, 112] were performed in the weak inversion regime, while we are interested in strong inversion. Actually, in strong inversion matters are somewhat simplified by not making this approximation. Doping fluctuations cause an equal, but opposite change in Q_D and Q_i, i.e. their contribution can be neglected. Gate depletion then causes parameter fluctuations through two mechanisms, namely 1) the increase in oxide thickness itself, and 2) the variation in this increase.
The increase in oxide thickness was modeled in subsection 4.1.3. The increase in threshold voltage mismatch it causes is calculated by replacing t_{ox} by $t_{ox} + t_{GD}$ in (4.79) and (4.80). Note that t_{GD} is a function of the gate bias. Therefore, the increase both affects threshold-voltage and current-factor mismatch in the simple model, developed in chapter 2 and used in chapter 3 to extract parameters.
Microscopic fluctuations in t_{GD} itself can be caused by fluctuations in t_{ox} and N_P. In modern-day CMOS technologies the oxide thickness is very well controlled. Therefore, it is not expected to give a significant contribution[24]. It follows from (4.34) that:

$$\Delta t_{GD} = \frac{-1}{t_{ox} + 2t_{GD}} \frac{\epsilon_{ox}^2 (V_{GS} - \phi_{MS} - \phi_B)}{\epsilon_{si} q N_P} \frac{\Delta N_P}{N_P}. \tag{4.85}$$

Modeling $\sigma_{\Delta N_P/N_P}$ is quite complicated. In general we can write:

$$\sigma^2_{\Delta N_P/N_P} = \frac{2\epsilon_{ox}}{\epsilon_{si} W L t_{GD} N_P} + \frac{A^2_{0,\Delta N_P/N_P,poly.str.}}{WL}. \tag{4.86}$$

The first term on the right-hand side is related to the number fluctuations of N_P and provides the lower boundary to the variation. For $N_P = 5 \cdot 10^{19}$ cm^{-3}, $t_{ox} = 1.5$ nm and $V_{GS}@t_{GD} = 0.75$ nm it is equal to $(0.4\ \%\mu\text{m})^2$. The second term models the increase in N_P related to the stochastic nature of the poly-silicon gate material. In subsection 4.3.6

[24]This conclusion will be further justified in subsection 4.3.5. However, note that it might change once high-k dielectrics are introduced.

it will be found that $A_{0,\Delta N_P/N_P,poly.str.} \approx 2.6$ %μm for the technology under consideration. $A_{0,\Delta N_P/N_P,poly.str.}$ is expected to decrease with decreasing grain size and is a function of e.g. the implantation conditions of the poly doping and the subsequent annealing steps.
The impact of this variation on the drain current follows from (4.57) and (4.65):

$$\left.\frac{\Delta I_D}{I_D}\right|_{\Delta t_{GD}} = -\left(\frac{1}{t_{ox}+t_{GD}} + \frac{|Q_D|}{\epsilon_{ox}(V_{GS}-V_T)}\right)\Delta t_{GD}. \qquad (4.87)$$

The terms in between the brackets are related to the current-factor and threshold-voltage dependence on t_{GD}, respectively. As example we fill in the same parameter values as earlier and $N_A = 1 \cdot 10^{18}$ cm^{-3}. By only taking number fluctuations in N_P into account, the first term of (4.87) gives a contribution of 0.10 %μm. The contribution of the second term is 0.40 mVμm. However, their importance increases for significant $A_{0,\Delta N_P/N_P,poly.str.}$. As mentioned earlier, t_{GD} is a function of the gate bias.
To end this subsection, note that effects related to gate depletion are expected to disappear once metal gates are introduced, as is planned for the 45 nm technology node.

4.3.4 Quantummechanical effects

To describe quantummechanical effects on threshold voltage fluctuations, usually three dimensional simulations are applied [79, 83, 84]. The use of a one dimensional approach was validated by simulations in [90]. In [24], a more analytical approach is followed, but numerical solving was required to obtain final results. In [83], 3D simulation results are compared to calculations which only take the quantummechanical increase in oxide thickness into account. In this subsection simple analytical expressions are obtained by extending the analysis of subsection 4.1.4 to take parameter fluctuations into account. Quantummechanical effects result in an increase in surface potential and an increase in oxide thickness. These effects will be dealt with separately.

Increase in surface potential. The quantummechanical increase in surface potential (see (4.35)) enhances threshold voltage fluctuations through two mechanisms: 1) It increases the depletion layer width, which results in extra fluctuations in the depletion layer charge (see subsection 4.3.1), and 2) the increase in ψ_s itself is proportional to Q_D and will therefore vary from transistor to transistor.
From (4.6), the quantummechanical increase in depletion layer width

(ΔW_D^{QM}) is calculated to be:

$$\Delta W_D^{QM} \cong \frac{\epsilon_{si}\Delta\psi_s^{QM}}{qN_A(W_D)W_D}. \tag{4.88}$$

The increase in $\sigma^2_{\Delta V_T}$ due to variations in Q_D follows immediately by replacing W_D with $W_D + \Delta W_D^{QM}$ in (4.79) and (4.80). Now look at the threshold-voltage variation due to the dependence of $\Delta\psi_s^{QM}$ on Q_D. As in subsection 4.3.2, consider a charge sheet with thickness dy at a distance y from the interface that has a doping concentration that is $\delta N_A(y)$ higher than the average N_A. The related decrease in depletion-layer width is smaller than classically expected, because of the quantummechanical increase in surface potential. Using (4.35) and (4.79), we can write:

$$\sigma^2_{\Delta' V_T} = \frac{q^2(t_{ox} + (\epsilon_{ox}/\epsilon_{si})d_{QM})^2}{WL\epsilon_{ox}^2} \cdot \tag{4.89}$$

$$\cdot \int_0^{W_D^{QM}} \left(1 - \frac{y - d_{QM}}{W_D^{QM} - d_{QM}}\right)^2 N_A(y)dy,$$

$$d_{QM} = \frac{2\epsilon_{si}^{1/3}B_{QM1}}{3|Q_D|^{1/3}} - \frac{\epsilon_{si}kT}{q|Q_D|}. \tag{4.90}$$

Figure 4.21 compares the calculated increase in $\sigma^2_{\Delta V_T}$ $(= 2\sigma^2_{\Delta' V_{GS}})$ due to quantum-mechanical effects with the results obtained in [83] by 3D atomistic simulations. Also shown is the calculated increase in $\sigma^2_{\Delta V_T}$ when a quantum-mechanical increase in oxide thickness of 0.37 nm is assumed, as was done in [83]. It is observed that both models give a reasonably accurate description of the simulation results.
In case of significant gate depletion, in (4.89) t_{ox} needs to be replaced by $t_{ox} + t_{GD}$. It follows that this somewhat reduces the relative increase due to quantummechanical effects.

Increase in oxide thickness. The increase in oxide thickness due to the non-zero peak location of the electron concentration is given by (4.36). Note again, that this increase only affects the current-factor. The threshold voltage shift is fully modeled by the increase in surface potential. It was found that $t_{QM} \propto E_{eff}^{-1/3}$. It follows that $\Delta t_{QM}/t_{QM} = -\Delta E_{eff}/3E_{eff}$. The mismatch in the drain-current is then given by:

$$\frac{\Delta I_D}{I_D} = \frac{-\Delta t_{QM}}{t_{ox} + t_{GD} + t_{QM}} = \frac{t_{QM}}{t_{ox} + t_{GD} + t_{QM}} \frac{\Delta E_{eff}}{3E_{eff}}. \tag{4.91}$$

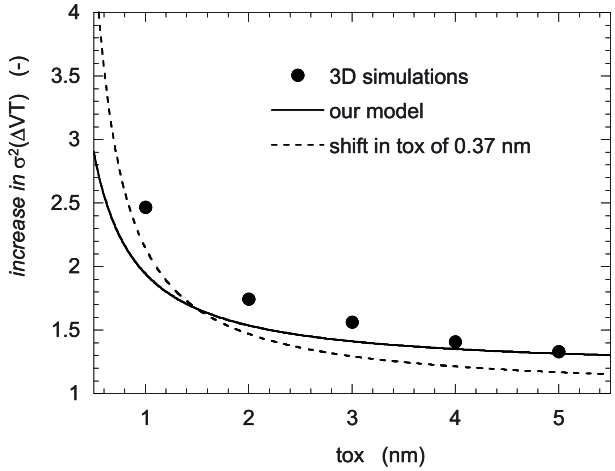


Figure 4.21. Increase in $\sigma^2_{\Delta V_T}$ due to quantummechanical effects. Compared are data obtained from 3D atomistic simulations (symbols) (manually extracted from [83]), our model (solid line) and the calculated results assuming an increase in oxide thickness of 0.37 nm (dashed line). $W = L = 50$ nm and $N_A = 5 \cdot 10^{18}$ cm^{-3}.

The mismatch in the effective field is calculated to be

$$\frac{\Delta E_{eff}}{E_{eff}} = \left.\frac{\Delta E_{eff}}{E_{eff}}\right|_{\Delta Q_D} + \left.\frac{\Delta E_{eff}}{E_{eff}}\right|_{\Delta t_{GD}} = \tag{4.92}$$

$$= \frac{(1-\eta)\Delta Q_D}{|Q_D + \eta Q_i|} - \frac{\eta \epsilon_{ox}(V_{GS} - \phi_{MS} - \phi_B)\Delta t_{GD}}{|Q_D + \eta Q_i|(t_{ox} + t_{GD} + t_{QM})^2}.$$

The first term on the right-hand side is related to doping fluctuations. This effect is introduced in (4.89) by replacing $(t_{ox} + (\epsilon_{ox}/\epsilon_{si})d_{QM})$ with $(t_{ox} + (\epsilon_{ox}/\epsilon_{si})d_{QM} + t_{QMQD})$, where:

$$t_{QMQD} = -\frac{(1-\eta)t_{QM}\epsilon_{ox}(V_{GS} - V_T)}{3(t_{ox} + t_{GD} + t_{QM})|Q_D + \eta Q_i|}. \tag{4.93}$$

At $N_A = 1 \cdot 10^{18}$ cm^{-3}, $t_{GD} = 0.75$ nm and $V_{GS} - V_T = 0.9$ V, $t_{QMQD} = -0.07$ nm, i.e. the effect can safely be neglected.
The second term on the right-hand side of (4.92) is related to fluctuations in gate depletion. This results in an extra term $+1/t_{QMGD}$ inside the brackets of (4.87). The thickness t_{QMGD} is given by:

$$t_{QMGD} = \frac{3|Q_D + \eta Q_i|(t_{ox} + t_{GD} + t_{QM})^3}{\eta \epsilon_{ox} t_{QM}(V_{GS} - \phi_{MS} - \phi_B)}. \tag{4.94}$$

Filling in the same numbers and $V_{GS} - \phi_{MS} - \phi_B = 1.2$ V, yields $1/t_{QMGD} \approx 1/80$ nm^{-1}, which is small compared to $1/(t_{ox} + t_{GD})$.

Finally note, that quantummechanical effects were not included in our treatment of gate-depletion in the previous subsection. To include them, in (4.85) and (4.87) t_{ox} needs to be replaced by $t_{ox} + t_{QM}$.

4.3.5 Mobility fluctuations

Variation in the current factor can be due to variation in width and length (this will be discussed in chapter 6), variation in oxide capacitance (see subsection 4.3.3) and variation in mobility. In this subsection mobility fluctuations will be discussed. Based on subsection 4.1.5, the responsible mechanisms are separated into fluctuations due to: 1) scattering to fixed oxide charges ($\Delta\mu_{fc}/\mu_{fc}$), 2) Coulomb scattering ($\Delta\mu_C/\mu_C$), and 3) surface-roughness scattering ($\Delta\mu_{sr}/\mu_{sr}$). From (4.22) and (4.37), the overall mismatch in drain current due to mobility mismatch is equal to:

$$\left.\frac{\Delta I_D}{I_D}\right|_{\Delta\mu} = \frac{\Delta\mu}{\mu} = \frac{\mu}{\mu_{fc}}\frac{\Delta\mu_{fc}}{\mu_{fc}} + \frac{\mu}{\mu_C}\frac{\Delta\mu_C}{\mu_C} + \frac{\mu}{\mu_{sr}}\frac{\Delta\mu_{sr}}{\mu_{sr}}. \tag{4.95}$$

Furthermore, extra fluctuations can arise from the variation in effective field. Also, fluctuations in the inversion-layer charge affect mobility through the screening terms in (4.40) and (4.43). We will proceed by modeling the different components[25].

Fluctuations in effective field. To calculate the impact of fluctuations in effective field on mobility, we write $\mu \propto E_{eff}^n$. From this it follows by definition that:

$$n \equiv \frac{\partial \ln(\mu)}{\partial \ln(E_{eff})}. \tag{4.96}$$

The magnitude of n lies in between 0 and -2, depending on the dominant scattering mechanisms, and it is a function of the applied bias conditions. When we assume that this bias dependence is not too strong, it follows that:

$$\left.\frac{\Delta\mu}{\mu}\right|_{\Delta E_{eff}} \cong n\frac{\Delta E_{eff}}{E_eff}. \tag{4.97}$$

The mismatch in the effective field is given by (4.92). In analogy with the definitions of t_{QMQD} and t_{QMGD} in the previous subsection, the

[25] A similar approach to the modeling of mobility fluctuations was presented in [65], but a somewhat different model was used as base. We distinguish ourselves at several points: 1) Scattering to fixed oxide charges is not neglected, since the oxide charge density of modern-day heavily-nitrided gate-oxides can be very high, 2) for Coulomb scattering, we take into account that the charge only scatters to a limited part of the channel doping, 3) we do not neglect fluctuations in screening by the inversion layer and 4) our model for fluctuations due to surface-roughness scattering is directly related to the physical properties of the oxide-silicon interface.

parameters $t_{\mu QD}$ and $t_{\mu GD}$ are defined as:

$$t_{\mu QD} = -n\frac{(1-\eta)\epsilon_{ox}(V_{GS}-V_T)}{|Q_D+\eta Q_i|}. \tag{4.98}$$

$$t_{\mu GD} = \frac{1}{n}\frac{|Q_D+\eta Q_i|(t_{ox}+t_{GD}+t_{QM})^2}{\eta\epsilon_{ox}(V_{GS}-\phi_{MS}-\phi_B)}. \tag{4.99}$$

Filling in the same parameter values as earlier and $n = -1/3$ yields[26] $t_{\mu QD} = 0.44$ nm and $t_{\mu GD} = -12$ nm. Both effects cannot be neglected and are introduced in (4.89) and (4.87) in the same way as t_{QMQD} and t_{QMGD}.

Scattering to fixed oxide charges. Fluctuations in the fixed oxide-charge were calculated in subsection 4.3.1. Introducing them in (4.39) to (4.41) gives:

$$\sigma^2_{\Delta\mu_{fc}/\mu_{fc}} = \frac{8.21\cdot 10^{-25}(z_\mu/|Q_i|)^{1/2}}{p_\mu^2}\frac{2}{N_f WL}. \tag{4.100}$$

These fluctuations are fully correlated with the threshold voltage fluctuations calculated with (4.80).

Coulomb scattering. Fluctuations in channel-doping cause fluctuations in the Coulomb-scattering-limited mobility (μ_C). According to (4.77), to estimate the magnitude, we need to know the sensitivity of the mobility to a variation in the doping concentration as a function of the depth. We assume that the range over which dopants impact the mobility is equal to the inversion layer thickness $z_\mu \approx 5-10$ nm, which is supported by the simulation results presented in [89]. It follows that:

$$\sigma^2_{\Delta\mu_C/\mu_C|\Delta N_A} \cong \frac{2}{WL\int_0^{z_\mu} N_A(y)dy}. \tag{4.101}$$

At a doping concentration of $N_A = 1\cdot 10^{18}$ cm^3, this results in $A_{0,\Delta\mu_C/\mu_C|\Delta N_A} \cong 1.8$ %μm. Note that remote impurity scattering is not taken into account in our formulation. More accurate expressions can be obtained by applying the theory out of [99, 100, 114], as was mentioned earlier.

The mismatch $\Delta\mu_C/\mu_C|_{\Delta N_A}$ is correlated with other equations related to doping fluctuations. Combining (4.89) and (4.101), the correlation is found to be:

[26] The value of $n = -1/3$ is related to phonon scattering.

$$\rho\left(\left.\frac{\Delta\mu_C}{\mu_C}\right|_{\Delta N_A},\Delta Q_D\right)\cong \tag{4.102}$$

$$\cong\frac{-\int_0^{z_\mu}\left(1-\frac{y-d_{QM}}{W_D^{QM}-d_{QM}}\right)N_A(y)dy}{\sqrt{\int_0^{z_\mu}N_A(y)dy\cdot\int_0^{W_D^{QM}}\left(1-\frac{y-d_{QM}}{W_D^{QM}-d_{QM}}\right)^2N_A(y)dy}},$$

For $N_A=1\cdot10^{18}$ cm^3 this equals $\rho\sim-50$ %.

Screening by the inversion layer. Besides fluctuations in doping, variations in μ_{fc} and μ_C are caused by fluctuations in the inversion-layer charge through the screening terms in (4.39) to (4.43). It follows that:

$$\left.\frac{\Delta\mu_{fc}}{\mu_{fc}}\right|_{\Delta Q_i}=\frac{2.27\cdot10^{-13}(z_\mu/|Q_i|)^{1/4}N_f}{p_\mu}\frac{\Delta Q_i}{Q_i}, \tag{4.103}$$

$$\left.\frac{\Delta\mu_C}{\mu_C}\right|_{\Delta Q_i}=\frac{\gamma_{BH}^4}{(1+\gamma_{BH}^2)^2\left(\ln(1+\gamma_{BH}^2)-\frac{\gamma_{BH}^2}{1+\gamma_{BH}^2}\right)}\frac{\Delta Q_i}{Q_i}. \tag{4.104}$$

Note that $\Delta Q_i/Q_i$ is equal to $\Delta I_D/I_D$ as calculated in the previous subsections, i.e. without taking mobility fluctuations into account. Combining this with (4.95) gives:

$$\left.\frac{\Delta I_D}{I_D}\right|_{\Delta Q_i}=\left(1+\frac{\mu}{\mu_{fc}}\left.\frac{\Delta\mu_{fc}}{\mu_{fc}}\right|_{\Delta Q_i}\frac{Q_i}{\Delta Q_i}+\frac{\mu}{\mu_C}\left.\frac{\Delta\mu_C}{\mu_C}\right|_{\Delta Q_i}\frac{Q_i}{\Delta Q_i}\right)\frac{\Delta Q_i}{Q_i}. \tag{4.105}$$

The factor inside the brackets has values in the range of 1.0-1.2, depending on the applied bias conditions.

Surface-roughness scattering. To calculate the effect of surface-roughness scattering, we need to go a bit deeper into the model that was presented in subsection 4.1.5. Generally, the roughness of the surface is described by a first-order autoregressive-process with an autocovariance function (R(r)) [101]:

$$R(r)=\Delta^2e^{-r/L_\Delta}, \tag{4.106}$$

where $\Delta\sim0.2-0.4$ nm represents the magnitude of the roughness and $L_\Delta\sim1-3$ nm its correlation length. This L_Δ can be related to the correlation length ($l_{\delta\mu_{sr}}$) as defined in subsection 4.2.1. Since the power-density function is the Fourier transform of the autocovariance function:

$$l_{\delta\mu_{sr}}^2=\int_0^\infty2\pi re^{-r/L_\Delta}dr=2\pi L_\Delta^2. \tag{4.107}$$

It is assumed that Δ^2 represents the variance of a process with a Gaussian distribution. In a device with area WL, Δ^2 is determined by $WL/l^2_{\delta\mu_{sr}}$ independent events. This results in $\sigma^2_{\Delta^2}/\Delta^2 = 2l^2_{\delta\mu_{sr}}/WL$ and:

$$\sigma^2_{\Delta\mu_{sr}/\mu_{sr}} = \frac{8\pi L^2_\Delta}{WL}. \tag{4.108}$$

For $L_\Delta = 2$ nm this gives $A_{0,\Delta\mu_{sr}/\mu_{sr}} = 1.0$ %μm. The expected linear dependence of the standard deviation on L_Δ, within reasonable limits, also followed from the simulations published in [82, 115]. The quantum-mechanics included in this work have been neglected in our model.
Note that from (4.106) we can now also calculate the variation in the oxide thickness itself. This is given by $A_{0,\Delta t_{ox}} = \sqrt{4\pi}\Delta \cdot L_\Delta = 0.7 - 4 \cdot 10^{-3}$ nmμm. As was concluded in subsection 4.3.3 this can be neglected. However, with further down scaling of the oxide thickness it might become an issue, and it would therefore be an interesting topic for future study.

4.3.6 Combination of all effects and comparison with experiments

In this subsection, the developed theory will be compared to experimental data. The 0.13 μm technology under consideration has a nominal gate length of 0.13 μm, a physical oxide-thickness of 2.0 nm, and a supply voltage of 1.5 V. Measured device pairs have $W = L = 1.0$ μm and are n-type. The drain bias is put to 50 mV. The channel doping profile is depicted in figure 4.20. It will be shown in section 5.3 that halos can be implanted through the gate and thus affect long-channel transistor behavior. This seriously complicates matters. Therefore, the transistors to which our models are experimentally compared did not receive a halo implantation. The magnitude of the different mobility components and the amount of fixed oxide-charge was determined in subsection 4.1.5. The doping-concentration in the gate is assumed to be $5{\cdot}10^{19}$ cm^{-3}. Only two unknown parameters are left, namely the increase in gate-doping fluctuations due to the poly-silicon structure of the gate ($A_{0,\Delta N_P/N_P,poly.str.}$) and the correlation length of the surface roughness (L_Δ). The latter is put to its minimal value of $L_\Delta = 1$ nm, which resulted in the best description of the experimental data. $A_{0,\Delta N_P/N_P,poly.str.}$ is obtained from a fit to this data.
In this subsection, we choose not to represent our data as $\Delta I_D/I_D$, but as:

$$\Delta U_T \equiv -(V_{GS} - V_T)\frac{\Delta I_D}{I_D}, \tag{4.109}$$

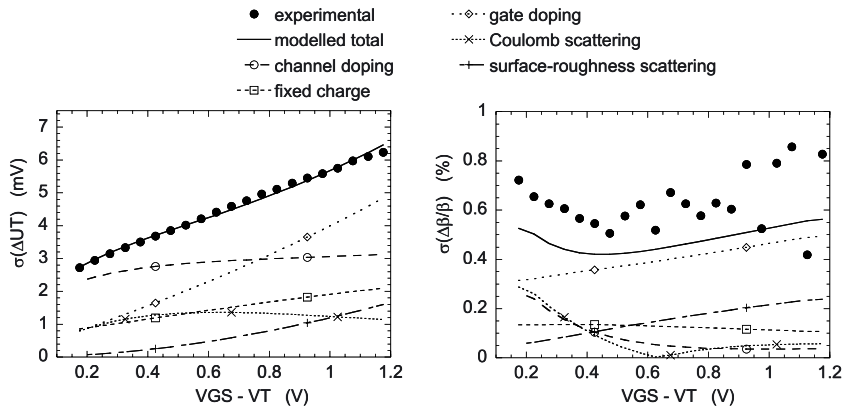


Figure 4.22. Modeled and experimental $\sigma_{\Delta U_T}$ and $\sigma_{\Delta\beta/\beta}$ as a function of the gate overdrive. Also shown are the contributions of the different fluctuation mechanisms that cause the mismatch. The parameters ΔU_T and $\Delta\beta/\beta$ are defined in (4.109) and (4.111), respectively.

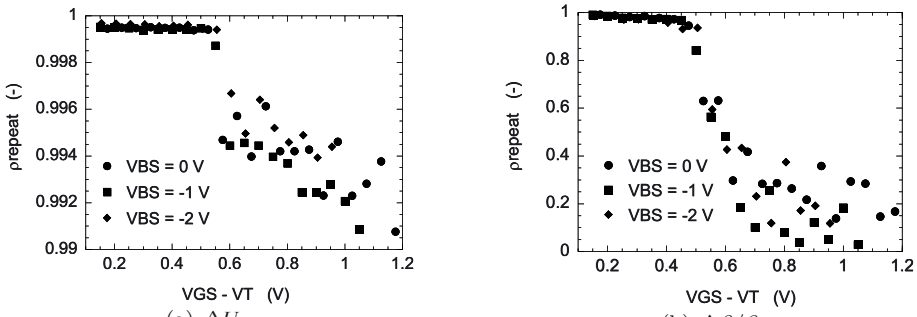
which is a measure of the difference in gate voltage at constant drain current, as follows from the following approximation:

$$\frac{\Delta I_D}{I_D} \cong \frac{-\Delta V_T}{V_{GS} - V_T} + \frac{\Delta\beta}{\beta}. \tag{4.110}$$

Based on this, the mismatch in the current factor is defined as

$$\frac{\Delta\beta}{\beta} \equiv \frac{-d\Delta U_T}{dV_{GS}}, \tag{4.111}$$

which is a measure of the mismatch in the transconductance. Analyzing ΔU_T instead of $\Delta I_D/I_D$ has the advantage of avoiding the singularity at $V_{GS} = V_T$. It is furthermore easier to relate a plot of ΔU_T versus V_{GS} to a mismatch in the threshold voltage and a mismatch in the current factor. It follows from (4.110) that $\Delta U_T = \Delta V_T - (V_{GS} - V_T)\Delta\beta/\beta$. In other words, when the tangent is taken at an overdrive voltage V_{ov}, its slope represents the mismatch in the current factor times -1 and the intercept with the y-axis is equal to the mismatch in the threshold voltage. The gate overdrive usually lies in the range of $V_{ov} = 0.15 - 0.4$ V. Figure 4.22 shows the experimental and modeled values of $\sigma_{\Delta U_T}$ and $\sigma_{\Delta\beta/\beta}$ as a function of the gate overdrive. The fit is performed on the $\sigma_{\Delta U_T} - (V_{GS} - V_T)$ curve. It is found that $A_{0,\Delta N_P/N_P,poly.str.} = 2.6$ %μm, which is a reasonable value. Based on this, it is seen that the magnitude of the $\sigma_{\Delta\beta/\beta} - (V_{GS} - V_T)$ curve is somewhat underestimated. The experimental curve displays a lot of scatter at higher values for the gate



(a) ΔU_T (b) $\Delta\beta/\beta$

Figure 4.23. Measurement repeatability of ΔU_T (a) and $\Delta\beta/\beta$ (b) as function of the gate overdrive. $V_{BS} = 0$ V.

bias. Therefore, no claims can be made about the functional shape. Measurement repeatability is plotted in figure 4.23. It is observed that the measurements of ΔU_T are almost fully repeatable, but that the repeatability drops slightly at higher gate bias. This is related to the switching of measurement range, as was earlier observed in subsection 3.3.2. The effect on the measurement of $\Delta\beta/\beta$ is much larger and it has been compensated for in figure 4.22.

Lets return to this figure. Also plotted are the contributions of the different components. These components originate from uncorrelated physical mechanisms and the modeled total is calculated by adding them quadratically. It is observed that all components play a role. For $\sigma_{\Delta U_T}$, at low gate overdrive, fluctuations in channel doping dominate. At higher gate bias, fluctuations in gate doping play a more prominent role. For $\sigma_{\Delta\beta/\beta}$, fluctuations in gate doping affect the mismatch over the whole bias range. At low gate overdrive also fluctuations in channel doping and Coulomb scattering are observed.

Most of the components in figure 4.22 consist of several subcomponents, as derived in the previous subsections. Figure 4.24 shows how the major components are built up. Subcomponent add up linearly, because they are caused by the same physical mechanism. In case of ΔU_T, it is observed that about 50 % of the total doping fluctuations is explained by the calculations presented in subsections 4.3.2 and 4.3.4. Furthermore, a big portion is caused by Coulomb scattering. The fluctuations due to Coulomb scattering were split up in a part that is fully correlated with ΔQ_D (as depicted in figure 4.24a) and a fully uncorrelated part (as depicted in figure 4.22). In case of $\Delta\beta/\beta$ (see figure 4.24b), besides the effects calculated in subsections 4.3.2 and 4.3.4, all effects play a role. At low gate bias, Coulomb scattering is the dominating subcomponent.

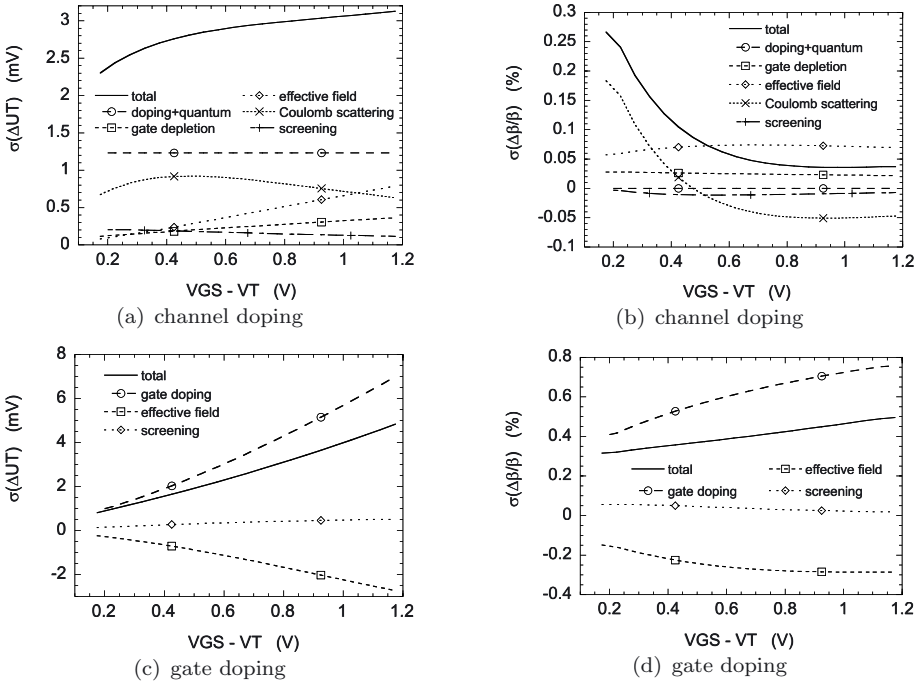


Figure 4.24. Calculation of major components out of figure 4.24 and their subcomponents

In figures 4.24)c+d it is observed that the fluctuations in gate-doping are partly compensated by the fluctuations in effective field. A somewhat thicker oxide reduces the inversion-layer charge, but it increases the mobility.

To further test the theory, the results out of figure 4.22 are extrapolated to lower values of the bulk bias, as is displayed in figure 4.25. For $\sigma_{\Delta U_T}$, the bulk bias dependence is well predicted at low values of the gate overdrive. However, the prediction is incorrect at higher $V_{GS} - V_T$. This might indicate (4.86) to be too simple to accurately describe fluctuations in gate doping. E.g. one can imagine that $A_{0,\Delta N_P/N_P,poly.str.}$ decreases with increasing t_{GD}. Another possibility is that our mobility model is not accurate enough. The bulk bias dependence of $\sigma_{\Delta\beta/\beta}$ is hidden by measurement noise.

Finally the correlation of ΔU_T with itself at different bias conditions is examined. This correlation shows to what extend ΔU_T is determined by the same physical mechanisms when the bias conditions are varied. Figure 4.26a plots the correlation of ΔU_T at $V_{GS} - V_T = 0.25$ V with ΔU_T at other values of the gate bias. This correlation is seen to decrease with

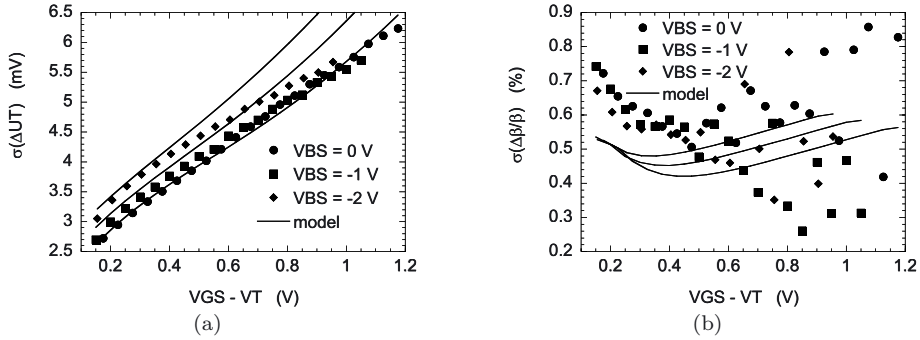


Figure 4.25. Modeled and experimental $\sigma_{\Delta U_T}$ (a) and $\sigma_{\Delta\beta/\beta}$ (b) as a function of the gate overdrive with the bulk bias as parameter.

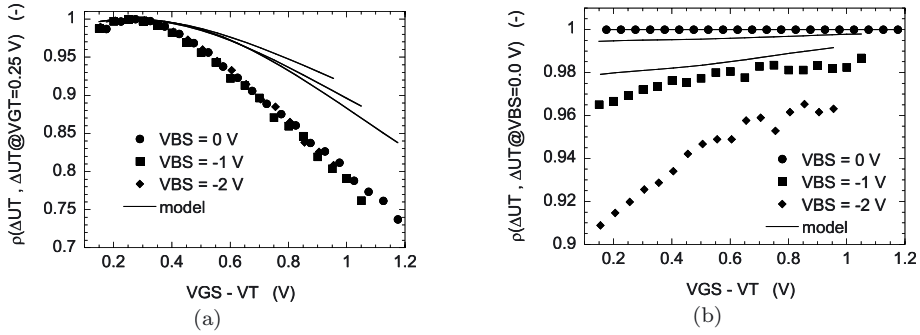


Figure 4.26. a) Correlation of ΔU_T at $V_{GS} - V_T = 0.25$ V with ΔU_T at other values of the gate overdrive as a function of this gate overdrive. b) Correlation of ΔU_T at $V_{BS} = 0$ V with ΔU_T at lower values of the bulk bias as a function of the gate overdrive.

increasing gate bias, which is reasonably well predicted by our models. Figure 4.26b shows the correlation of ΔU_T at $V_{BS} = 0$ V with ΔU_T at lower values of the bulk bias. The decrease in correlation with decreasing bulk bias is underestimated by our model, but the decrease itself is observed to be small.

4.3.7 Discussion

From the experimental work presented in the previous subsection it can be concluded that our physical models provide a good prediction of the order of magnitude of the matching properties of the MOSFET. In order to obtain a good description only two fitting parameters are required that are related to the magnitude of doping fluctuations in the gate and to the correlation length of the surface-roughness. We

also conclude that Coulomb scattering plays a more important role than it has been attributed in literature. Besides explaining the mismatch in current factor at low values of the gate bias, it also causes an apparent mismatch in the threshold voltage. The more or less constant contribution of the mismatch in Coulomb scattering to ΔU_T, as observed in figures 4.22 and 4.24a, can be explained as follows: To first order, due to screening by the inversion layer, the impact of Coulomb scattering is inversely proportional to the inversion layer charge. This gives rise to a $1/(V_{GS} - V_T)$ dependence, which means that fluctuations due to Coulomb scattering cannot be distinguished from fluctuations in the threshold voltage. Therefore, it could indeed explain the generally observed inconsistency between the calculated impact of doping fluctuations on the threshold voltage and the experimentally observed values. Although predicting correct orders of magnitude, our model only showed limited predictive properties when extrapolated to other bias conditions. This can have two origins: 1) The structural properties of the MOSFET are not fully known, especially for the poly-silicon gate. 2) The applied model for gate depletion is very simple, while the models for mobility are semi-empirical at best. Furthermore, we are extrapolating these models to regions for which they were not developed and in which they were not tested. E.g., our equations related to the fluctuations in Coulomb scattering predict that it cannot be neglected, even at high values of the gate bias (see figure 4.22). It would be worthwhile to try to develop these models, based on more sound physical principles, as was for example recently done for the surface potential [116]. Another approach would be to try to accurately simulate the above mentioned effects.
Note that the presented models do have their use. They allowed us to qualitatively identify dominant fluctuation mechanisms and at which bias conditions they play a role. The created insight is important when analyzing process splits, as will be found in the next chapter. It also suggest possible technology improvements or future bottlenecks. Fluctuations due to the poly-grain structure of the gate are process related and might be improved. Fluctuations due to Coulomb scattering can be reduced by reducing the doping concentration close to the oxide-silicon interface, e.g. by implementing an undoped epitaxially grown silicon layer on top of the doped substrate. Remote Coulomb scattering to dopants in the gate could be a possible bottleneck when oxide thicknesses are reduced further.

4.4 Conclusions

In this chapter the physical origins of MOSFET mismatch were examined. Firstly, a brief overview of MOSFET operation was presented.

Secondly, in the main part of the chapter, it was examined how microscopic mismatch affects macroscopic transistor behavior and which physical mechanisms are responsible for the microscopic mismatch.
It was found that the commonly used $\sigma_{\Delta P} \propto 1/\sqrt{WL}$ law only holds in the linear regime at low drain bias. In saturation a logarithmic deviation with the length has been observed. This deviation is caused by the higher resistivity of the channel at the drain side than at the source side. Its magnitude is limited by the correlation length of the mismatch causing stochastic process or by short-channel effects. Short- or narrow-channel devices also show deviations, as was reported earlier in literature. For short transistors, these deviations are mainly caused by geometrical effects. For narrow transistors other effects like the lower doping level and sidewall roughness can play a role.
In weak inversion, several mechanisms have been identified that cause deviations from the $\sigma_{\Delta P} \propto 1/\sqrt{WL}$ law. Firstly, in weak inversion, local variations are large, because of the exponential dependence of the inversion-layer-charge density on the surface potential. The current tends to flow around regions with high resistivity, while it prefers to flow in regions with low resistivity. This effect has been investigated by simulating resistor networks. Secondly, halos and narrow-channel effects have a relatively large impact on the mismatch in weak inversion. A strong halo determines most of the weak-inversion current, even for reasonably long transistors. Therefore, it effectively decreases the channel length, which leads to a relative increase in the magnitude of the fluctuations. The narrow-channel effect gives rise to similar behavior. Most of the weak-inversion current flows in the edge transistors. For wider transistors this effectively decreases the width, which again results in a relative increase of the magnitude of the fluctuations. For NMOS transistors, this width effect was found to be the main reason for the difference in mismatch between weak and strong inversion. PMOS transistors were found not to possess severe narrow-channel effects, and do not suffer from this behavior.
Most of the above mentioned effects were expected to impact the symmetry of the MOSFET. At low drain bias the device was found to be fully symmetrical. In saturation, asymmetry was observed for long transistors, because of the higher resistivity of the channel at the drain side. For short transistors, channel-length modulation was seen to cause asymmetry. In weak inversion it was observed that higher drain bias lowers the halo barrier at the drain side of the transistor, again giving rise to a loss of symmetry.
The physical mechanisms that are responsible for microscopic fluctuations were identified as: 1) doping fluctuations in the channel, 2) doping

fluctuations in the gate, 3) fluctuations in fixed-oxide charge, 4) fluctuations due to Coulomb scattering and 5) fluctuations due to surface-roughness scattering. The fluctuations in gate doping can be enhanced by the poly-grain structure of the gate material. These mechanisms affect transistor behavior by influencing: 1) the threshold voltage, 2) the amount of gate depletion, 3) the magnitude of quantummechanical effects, 4) the effective field, 5) the amount of carrier screening and 6) the mobility in general. It was found that none of the above mentioned effects can safely be neglected. At low gate overdrive the most dominant mechanism was found to be the fluctuation in channel doping. Besides directly influencing the threshold voltage mismatch, it also causes an apparent mismatch in threshold voltage due to Coulomb scattering. At higher values of the gate bias fluctuations in gate doping become more prominent.

The predictive quality of the developed model was tested by varying the bulk bias and by looking at the correlation of the mismatch at different bias conditions. The bulk bias dependence was found to be well predicted at low values of the gate overdrive voltage. At higher gate bias the prediction became less accurate. Correlations at different bias conditions were reasonably well predicted. To obtain higher accuracy, it was reasoned that the models related to variations in gate doping and the models related to the variation in mobility need to be improved. The presented equations provide qualitative insight and can be used in the optimization of a technology with respect to its matching performance.

Chapter 5

TECHNOLOGICAL ASPECTS

Until now, we have addressed the matching properties of MOS transistors from a device point of view. However, these devices need to be fabricated. Ideally, the fabrication process does not influence the matching behavior of a technology. In this case, the magnitude of the microscopic fluctuations is lower bounded by the Poisson statistics attributed to the dopants, as was discussed in the previous chapter. However, we will see that the fabrication process can have an impact on the matching behavior. This can be due to the intrinsic properties of a certain process step or of a material used in this step, but it can also be caused by unwanted side effects.

In literature, only a limited amount of papers have been published about the impact of processing on MOS transistor mismatch. These deal with the impact of the granular structure of the gate material [21, 113], channel engineering [10] and the impact of charging damage during processing [117]. Others deal with layout related issues like the impact of the proximity of a capacitor [47], the impact of metal lines [20, 46] or more complex layouts of the transistors themselves [118, 119]. In [8] an overview is presented of how the matching performance of a 0.18 μm CMOS technology improves with the optimization of several process steps. However, it is not specified how these optimizations are done.

In this chapter it is not attempted to present a full overview of the impact of CMOS process steps on MOSFET mismatch. We shall limit ourselves to the examples encountered during this work. These will demonstrate how certain process steps can have a devastating effect on the matching properties of a technology, while hardly affecting average MOSFET operation. In the first section of this chapter the examined technologies are briefly introduced. The second section examines the choice of gate

material and the third section looks at the impact of the halo implantation. In the fourth section the examined technologies are compared and the impact of scaling on the matching properties is discussed. The fifth section briefly addresses the matching behavior for future device architectures after which this chapter is concluded.

5.1 Technology descriptions

In this chapter we will examine four CMOS technologies that were developed in IMEC. These technologies are optimized towards transistors with physical gate lengths of 100 nm to 180 nm. A schematic overview of their process flows is presented in table 5.1. The different steps will now briefly be described.

Shallow Trench Isolation (STI). Shallow trench isolation is applied to electrically isolate one device from the other. This is achieved by depositing oxide in etched trenches. The depth of these trenches ranges from 325 nm to 400 nm for the examined technologies.

Deep well implantations. The deep well implantation is also used to provide isolation between transistors. This implantation fixes the doping concentration under the STI. Together with the thickness of the STI, this doping concentration determines the possible leakage from one transistor to the other: The higher the doping level, the better the isolation. The n-well is implanted with phosphorus with an energy of 380 keV and a dose of $1.0 \cdot 10^{13}$ cm^{-2}. The p-well is implanted with boron with an energy of 180 keV and a dose of $1.2 \cdot 10^{13}$ cm^{-2}.

Channel implantations. The channel implantation is performed at a lower energy than the deep well implantation. It determines the doping concentration at the top silicon interface and by this the threshold voltage. In the n-well an arsenic implantation is used with energies in the range of $100 - 200$ keV and doses in the range of $3 \cdot 10^{12} - 6 \cdot 10^{12}$ cm^{-2}. In the p-well a boron implantation is used with energies in the range of $20 - 40$ keV and doses in the range of $0.5 \cdot 10^{13} - 2 \cdot 10^{13}$ cm^{-2}.

Gate stack. The gate stack consists of the gate insulator (SiO_2) and the gate electrode. Nitrogen is introduced in the gate oxide to prevent boron penetration for PMOS transistors. In case of the $L_{nominal} = 100$ nm and $L_{nominal} = 130$ nm technologies, the very high nitrogen concentration also gives rise to an increase in the relative permittivity to a value in between $\epsilon_r \approx 3.9$ and $\epsilon_r \approx 6$, thus reducing the SiO_2-equivalent oxide thickness. These equivalent thicknesses are also listed in table 5.1 for the examined technologies.

For the $L_{nominal} = 180$ nm technology the gate consists of amorphous silicon, that after recrystallization results in grain sizes of ~ 100 nm. Polysilicon is used for technologies with shorter nominal gate lengths. This

Table 5.1. Front end of line process steps of the four examined CMOS technologies.

$L_{nominal}$ (nm)	100	130	150	180
V_{DD} (V)	1.2	1.5	1.5	1.8
oxide thickness (nm)	1.5	2.0	3.0	3.5
gate pre-doping (nm)	yes	no	no	no
Shallow Trench Isolation				
Deep NWELL implantation				
NWELL channel implantation				
Deep PWELL implantation				
PWELL channel implantation				
Gate stack				
N-halo implantation				
N-LDD implantation				
P-halo implantation	yes	yes	no	no
P-LDD implantation	yes	yes	no	no
Spacers				
N-HDD and P-HDD implantations				
Silicidation				

material has grain sizes of ~ 30 nm. In case of the $L_{nominal} = 100$ nm technology, gate pre-doping is applied for the NMOSFETs to lower the gate resistance and to decrease the gate depletion. This gate pre-doping consists of a phosphorus implantation with an energy of 25 keV and a dose of $2.0 \cdot 10^{15}$ cm^{-2}.

Together with the channel doping, the gate stack determines the long-channel threshold voltage and current factor. These parameters are shown in figure 5.1 as a function of the gate length[1]. As expected, the threshold voltage decreases and the current factor increases with decreasing oxide thickness.

Halo implantations. The halo implantations are performed to counter the short-channel effect. They are responsible for the roll-up of the $V_T - L$ curves in figure 5.1. In a well optimized technology the nominal transistor lies close to the peak of this curve. This is observed to be the case for the NMOS transistors. The PMOS transistors are seen to be

[1] In this chapter, the average values of the threshold voltages and current factors are extracted with the maximum slope method. However, for reasons explained in chapter 3, the mismatch in these parameters is extracted by the three-points method.

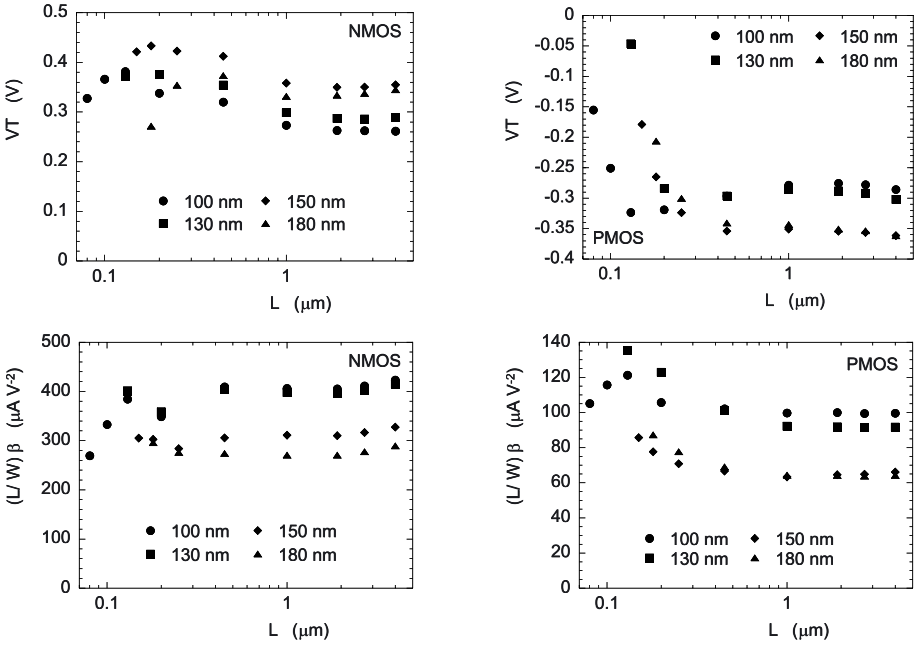


Figure 5.1. Threshold voltage and normalized current factor as a function of the gate length for the four CMOS technologies that are examined in this chapter. The legends list the nominal gate lengths of the technologies.

not as well optimized.

For the halos of the NMOS transistors a boron or BF_2 implantation is used at an angle in the range of $25° - 45°$, with energies in the range of $10 - 75$ keV and with doses in the range of $1 \cdot 10^{13} - 3 \cdot 10^{13}$ cm^{-2}. For the halos of the PMOS transistors an arsenic implantation is used at an angle in the range of $25° - 45°$, with energies of ~ 100 keV and with doses of $\sim 2 \cdot 10^{13}$ cm^{-2}. Note that for the $L_{nominal} = 150$ nm and $L_{nominal} = 180$ nm technologies the halo implantations for the PMOSFETs were not yet introduced.

Lightly Doped Drain (LDD) implantations, spacers and Highly Doped Drain (HDD) implantations. Two implantation steps are executed in order to dope the source and drain regions. Firstly, the low energy LDD implant is performed. This determines the junction depth close to the channel, which impacts short-channel behavior. However, the resistance of the LDD regions is not negligible and it decreases the drive current. The n-LDD is implanted with arsenic with energies in the range of $5 - 30$ keV and doses in the range of $1 \cdot 10^{14} - 2 \cdot 10^{15}$ cm^{-2}. The p-LDD is implanted with boron or BF_2 with energies in the range of

$1 - 10$ keV and doses in the range of $1 \cdot 10^{14} - 2 \cdot 10^{15}$ cm^{-2}.
Secondly, after forming spacers at the side of the gate, the HDD is implanted at a higher energy and with a higher dose than the LDD implantation. This lowers the resistance of the main part of the source and drain regions, but because of the spacers it doesn't affect the junction depth and the doping level close to the channel. Furthermore, the HDD implantation also dopes the gate. The n-HDD is implanted with arsenic with energies in the range of $25 - 75$ keV and doses in the range of $1 \cdot 10^{15} - 1 \cdot 10^{16}$ cm^{-2}. The p-HDD is implanted with boron or BF_2 with energies in the range of $2 - 25$ keV and doses in the range of $2 \cdot 10^{15} - 5 \cdot 10^{15}$ cm^{-2}. A schematic overview of the resulting doping profile was presented in figure 4.1. Note that for the $L_{nominal} = 150$ nm and $L_{nominal} = 180$ nm technologies the LDD implantations for the PMOSFETs were omitted, because of a too strong lateral diffusion of the HDD regions.

Thermal steps. After most implantation steps a thermal step is given in order to electrically activate the dopants. These thermal steps also cause diffusion of dopants from strongly doped regions to regions with lower doping levels.

Silicidation. To further lower the resistance of the gate, source and drain, a titanium-cobalt silicide is formed on top of these regions.

5.2 Impact of the gate

As was reasoned in the previous chapter, the gate can influence the matching properties of a technology 1) by increasing the effective oxide thickness due to gate depletion, 2) by fluctuations in the gate doping itself and 3) by boron penetration through the gate oxide in case of PMOS transistors [21]. In this section two experiments related to the processing of the gate stack are evaluated. The first investigates the impact of changing the gate material from amorphous silicon to polycrystalline silicon. The second looks at the way the gate is doped.

5.2.1 Amorphous or poly-crystalline silicon as gate material?

The impact of the gate material on transistor performance was examined in the $L_{nominal} = 150$ nm technology with $t_{ox} = 3.0$ nm. Figure 5.2 shows the average threshold voltage and normalized current factor for transistors processed with poly-silicon gate material and for transistors with amorphous gates. No significant differences are observed for the NMOS transistors. For the PMOS transistors a higher absolute

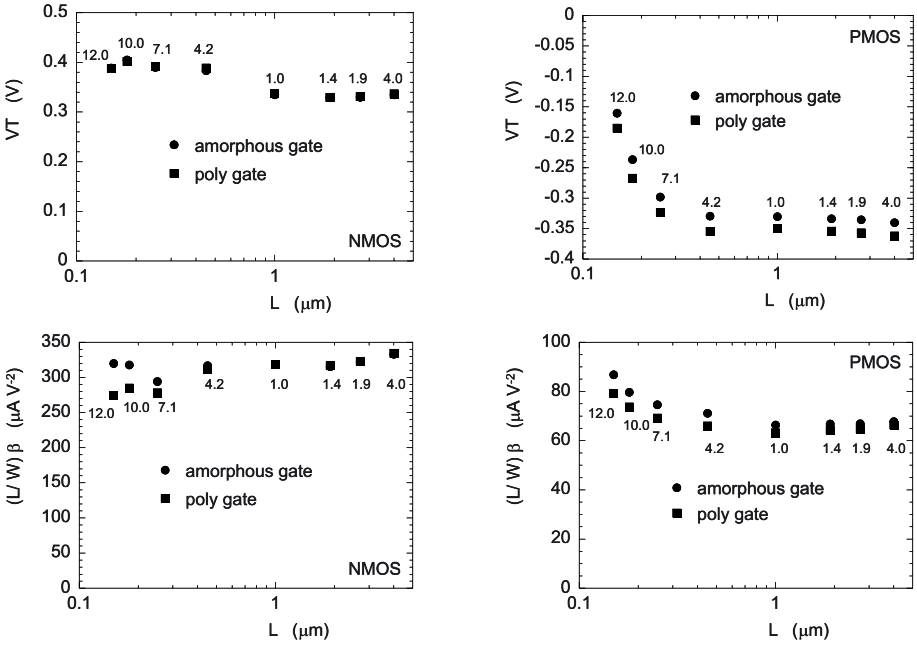


Figure 5.2. Threshold voltage and normalized current factor as a function of the gate length for the $L_{nominal} = 150$ nm technology with $t_{ox} = 3.0$ nm. Transistors with poly-silicon gates and with amorphous gates are compared. The numbers in the figures give the width of the transistors in μm.

Table 5.2. Proportionality constants of the $\sigma_{\Delta P} \propto 1/\sqrt{WL}$-law for transistors with an amorphous gate and with poly-silicon gate material. Also the one-sigma confidence intervals are given.

gate material	$A_{0,\Delta V_T}$ (mVμm)	$A_{0,\Delta\beta/\beta}$ (%μm)	$A_{0,\Delta V_T}$ (mVμm)	$A_{0,\Delta\beta/\beta}$ (%μm)
	NMOS		PMOS	
amorphous	5.41 ± 0.16	1.25 ± 0.15	5.98 ± 0.36	1.40 ± 0.14
poly	4.31 ± 0.33	1.01 ± 0.12	3.27 ± 0.37	1.09 ± 0.08

threshold voltage and a slightly lower current factor are observed for the devices with a poly-silicon gate. This means that these devices suffer more strongly from gate depletion than the devices with an amorphous gate. The shift in threshold voltage can also be partly caused by a higher level of boron penetration for the transistors with an amorphous gate.

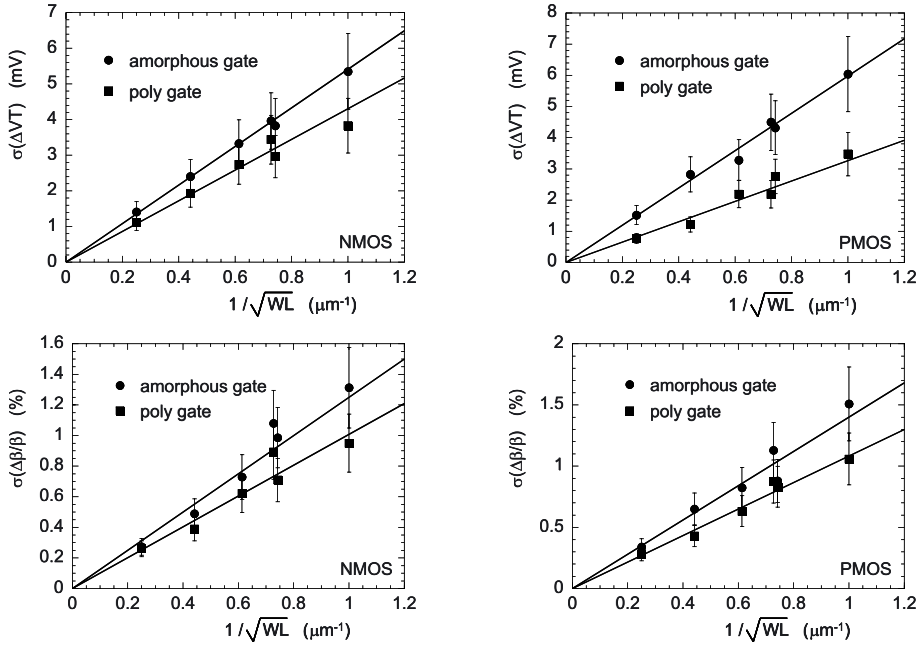


Figure 5.3. $\sigma_{\Delta V_T}$ and $\sigma_{\Delta\beta/\beta}$ as a function of $1/\sqrt{WL}$ for transistors with $W > 0.5$ μm and $L > 0.4$ μm. Transistors with poly-silicon gates and with amorphous gates are compared. Error bars represent 99 % confidence intervals.

However, if this effect would be dominant, this would cause an opposite shift in the current factor.

Figure 5.3 shows $\sigma_{\Delta V_T}$ and $\sigma_{\Delta\beta/\beta}$ as a function of $1/\sqrt{WL}$ for the transistors from table 2.1 with $W > 0.5$ μm and $L > 0.4$ μm. It is observed that transistors with poly-silicon gates possess a better matching performance than those with amorphous gates. Proportionality constants of the expected linear relationships are listed in table 5.2. The use of poly-silicon gate material results in better matching behavior because of the smaller grain size compared to the grain size of the amorphous gate after recrystallization. The correlation lengths of mismatch causing stochastic processes in the gate are directly proportional to this grain size. This also means that, even though the PMOS transistors with a poly-silicon gate suffer more from gate depletion, the impact of the variation in this gate depletion is smaller, since it is more effectively averaged out over the transistor. In general, we can state that one way to decrease the impact of the gate on the matching behavior of a technology is to decrease the poly-grain size of the gate material. An even better matching performance could be achieved after the introduction of metal gates.

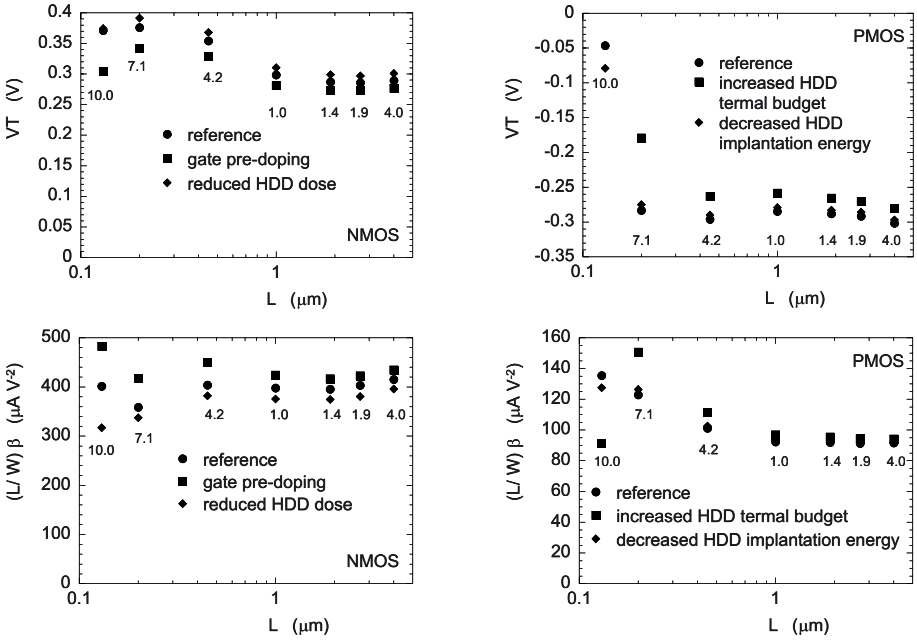


Figure 5.4. Threshold voltage and normalized current factor as a function of the gate length for the $L_{nominal} = 130$ nm technology with $t_{ox} = 2.0$ nm. The figures related to the NMOS transistors show data for the reference, for transistors with gate pre-doping and for transistors that received a reduced HDD dose. The figures related to the PMOS transistors show data for the reference, for transistors with increased thermal budget and for transistors for which the HDD implantation energy was decreased. The numbers in the figures give the width of the transistors in μm.

5.2.2 Impact of the gate doping

In [21] it was reasoned that the temperature of the rapid thermal anneal step after the HDD implantation impacts the matching behavior in the following way: When the temperature is too low, the matching performance is degraded by too strong gate depletion. However, when the temperature is too high, it is degraded, because of the boron penetration that occurs for PMOS transistors.

We investigated the impact of the gate doping on the $L_{nominal} = 130$ nm technology with $t_{ox} = 2.0$ nm. First consider the NMOS transistors. The following variations in the process were applied: For the reference, the arsenic HDD implantation is performed with an energy of 40 keV and a dose of $6.0 \cdot 10^{15}$ cm^{-2}. To reduce possible gate depletion, gate pre-doping was added to two device wafers. This gate pre-doping consists of a phosphorus implant with an energy of 25 keV and a dose of

Table 5.3. Proportionality constants of the $\sigma_{\Delta P} \propto 1/\sqrt{WL}$-law related to the measurements presented in figure 5.5. Also the one-sigma confidence intervals are given.

experimental split	$A_{0,\Delta V_T}$ (mVμm)	$A_{0,\Delta\beta/\beta}$ (%μm)
	NMOS	
reference	4.01 ± 0.32	1.04 ± 0.08
gate pre-doping	3.64 ± 0.33	1.06 ± 0.12
reduced HDD dose	4.29 ± 0.49	1.18 ± 0.10
	PMOS	
reference	2.51 ± 0.26	0.98 ± 0.10
increased HDD thermal budget	2.51 ± 0.28	0.89 ± 0.11
reduced HDD implantation energy	2.59 ± 0.29	0.95 ± 0.08

$2.0 \cdot 10^{15}$ cm^{-2}. The third NMOS process split had a reduced HDD dose of $4.0 \cdot 10^{15}$ cm^{-2} for which increased gate depletion is expected. For the PMOS reference the boron HDD is implanted with an energy of 4.0 keV and a dose of $3.0 \cdot 10^{15}$ cm^{-2}. For the second experimental split the thermal budget after the HDD implant is increased, which is expected to result in lower gate depletion, but possibly higher boron penetration. For the third split the HDD dose is increased to $4.5 \cdot 10^{15}$ cm^{-2}, which should also lead to lower gate depletion.

Figure 5.4 compares the average threshold voltage and normalized current factor for the examined experimental splits. It is indeed observed that in cases with expected lower gate depletion the absolute value of the threshold voltage is lower and the value of the current-factor is higher with respect to the reference. The opposite behavior is observed when the gate depletion is expected to be higher.

Figure 5.5 shows $\sigma_{\Delta V_T}$ and $\sigma_{\Delta\beta/\beta}$ as a function of $1/\sqrt{WL}$ for the same experimental splits. The proportionality constants are listed in table 5.3. In case of the PMOS devices, no significant differences are observed. For the NMOSFETs, the transistors that suffer most from gate depletion, also posses the worst matching characteristics. However, the differences are small and it is doubtful if they are significant. More accuracy might be obtained in a future experiment by increasing the population size above the 84 device pairs per dimension that were available for this experiment. However, at this point we can conclude that the gate doping is not the dominant mismatch causing effect of the examined technology.

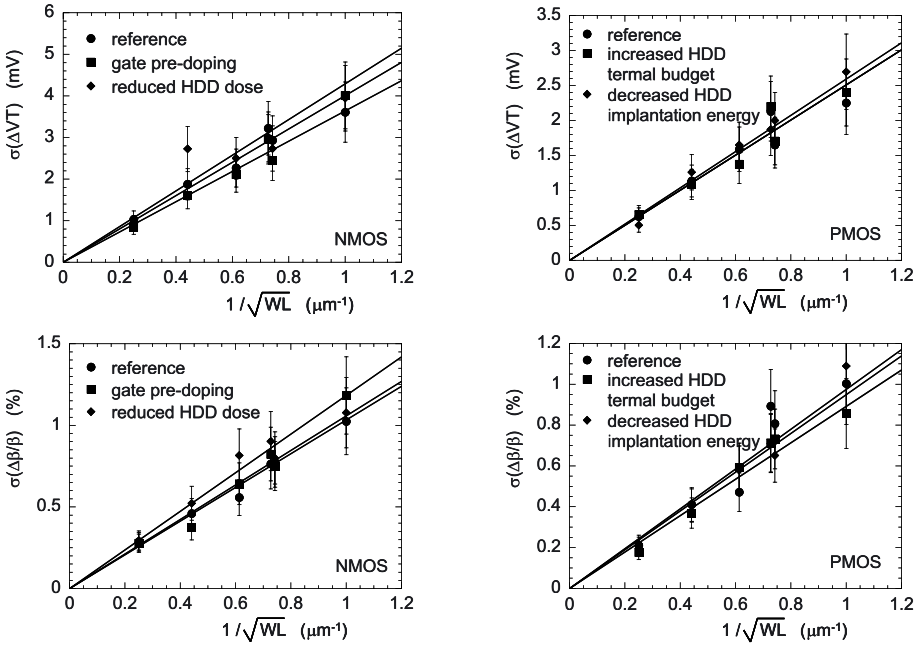


Figure 5.5. $\sigma_{\Delta V_T}$ and $\sigma_{\Delta\beta/\beta}$ as a function of $1/\sqrt{WL}$ for transistors with $W > 0.5$ μm and $L > 0.4$ μm. The experimental splits are the same as in figure 5.4. Error bars represent 99 % confidence intervals.

5.3 Impact of the halo implantation

The impact of the halo implantation on MOS transistor matching is examined on the $L_{nominal} = 130$ nm technology with $t_{ox} = 2.0$ nm by varying the implantation conditions as listed in table 5.4. Four experimental splits are investigated: 1) No halo is implanted, 2) the reference implantation conditions, 3) the halo dose is increased and 4) the implantation angle is increased.

Figure 5.6 shows the average threshold voltage and normalized current factor as a function of the gate length for the different experimental splits. As expected, the short-channel threshold voltage becomes larger for increasing halo dose. The same effect is observed when the implantation angle is increased, which is due to the fact that doping close to the source and drain regions contributes less to the threshold voltage than doping in the center region of the channel. For the current factor no significant differences are observed for the long-channel transistors. For short-channel transistors the current factor is largest when no halos are present. This is most probably due to the smaller effective channel length of these transistors. This effect is observed to be present in a

Table 5.4. Experimental splits on the halo implantation conditions

	no halo	reference	increased dose	increased angle
			NMOS	
type		BF_2	BF_2	BF_2
angle	no halo	35°	35°	45°
dose (cm^{-2})		$1.6 \cdot 10^{13}$	$2.5 \cdot 10^{13}$	$1.6 \cdot 10^{13}$
energy (keV)		120	120	120
			PMOS	
type		As	As	As
angle	no halo	35°	35°	45°
dose (cm^{-2})		$2.1 \cdot 10^{13}$	$3.0 \cdot 10^{13}$	$2.1 \cdot 10^{13}$
energy (keV)		120	120	120

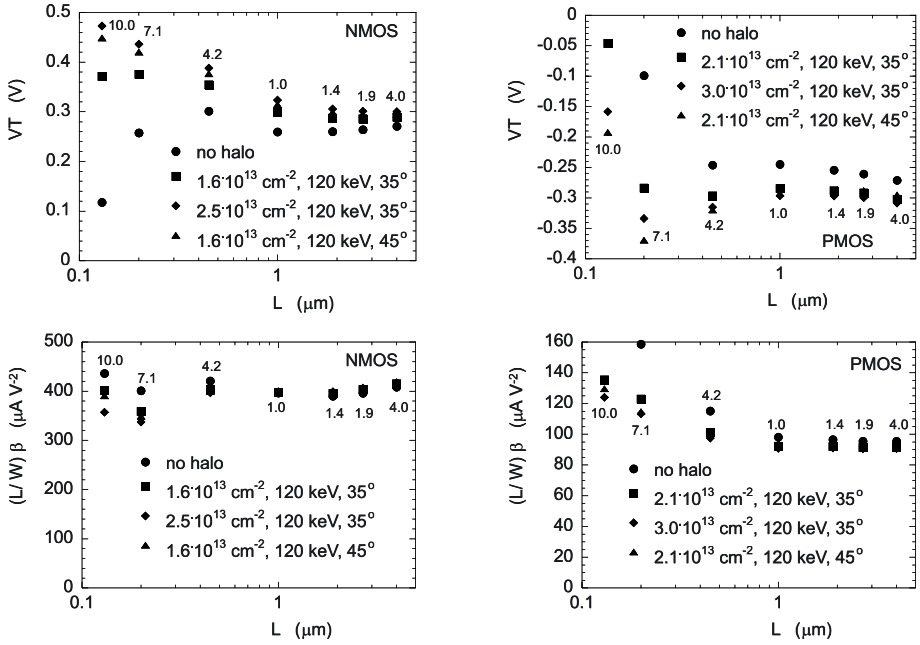


Figure 5.6. Threshold voltage and normalized current factor as a function of the gate length for the $L_{nominal} = 130$ nm technology with $t_{ox} = 2.0$ nm. The displayed experimental splits are described in table 5.4. The numbers in the figures give the width of the transistors in μm.

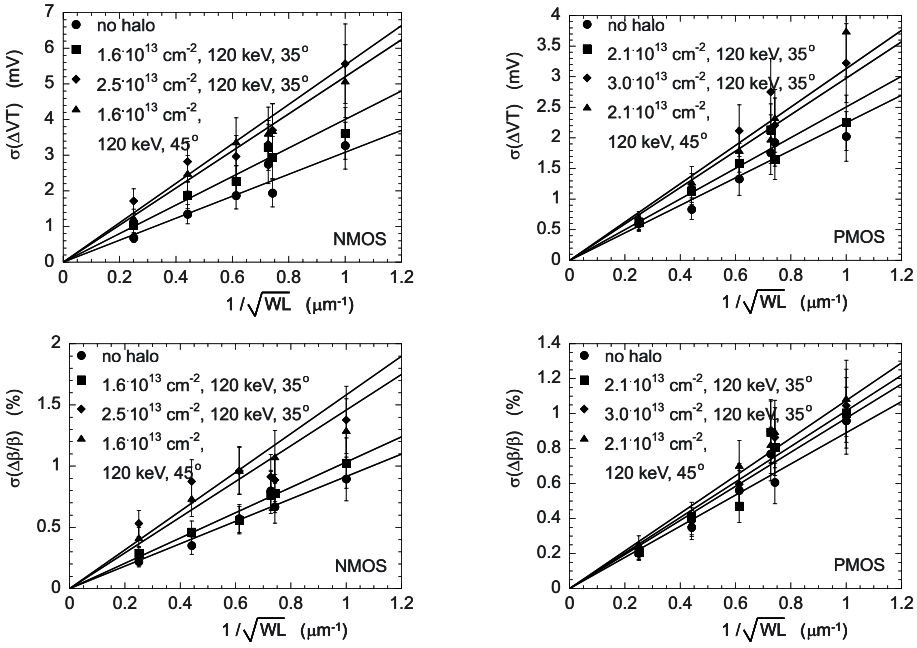


Figure 5.7. $\sigma_{\Delta V_T}$ and $\sigma_{\Delta\beta/\beta}$ as a function of $1/\sqrt{WL}$ for transistors with $W > 0.5$ μm and $L > 0.4$ μm. The displayed experimental splits are listed in table 5.4. Error bars represent 99 % confidence intervals.

much stronger way for the PMOS transistors than for the NMOS transistors. Also the extra channel doping introduced by the halos causes larger Coulomb scattering, which reduces the current factor.
We will now look at the extra parameter fluctuations that the halos introduce. The first subsection of this section discusses the matching properties for long and wide transistors, while the second subsection deals with short- and narrow-channel effects.

5.3.1 Long- and wide-channel transistors

Figure 5.7 shows $\sigma_{\Delta V_T}$ and $\sigma_{\Delta\beta/\beta}$ as a function of $1/\sqrt{WL}$ for the experimental splits described by table 5.4. The proportionality constants are listed in table 5.5. Since halos are only supposed to be located around the source and drain regions, they are not expected to affect the matching properties of long-channel transistors. However, it is observed that the mismatch increases when halos are implanted. Furthermore, it is observed that $\sigma_{\Delta V_T}$ and $\sigma_{\Delta\beta/\beta}$ are proportional to $1/\sqrt{WL}$ for all splits. This means that the mismatch causing stochastic process is not only located at the source and drain sides of the transistor, but that it

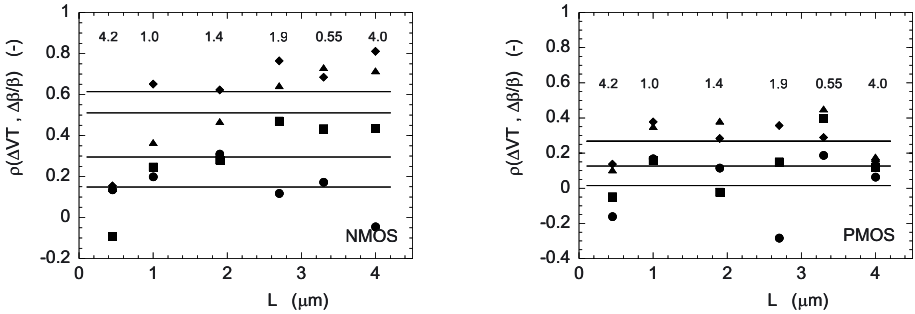


Figure 5.8. Correlation between ΔV_T and $\Delta\beta/\beta$ as a function of the gate length. The displayed experimental splits are listed in table 5.4. Symbols have the same meaning as in figure 5.7. The numbers in the figures give the width of the transistors in μm.

Table 5.5. Proportionality constants of the $\sigma_{\Delta P} \propto 1/\sqrt{WL}$-law and the long-channel correlation between ΔV_T and $\Delta\beta/\beta$ for transistors with different halo implantation conditions, as listed in table 5.4. Also the one-sigma confidence intervals are given.

	$A_{0,\Delta V_T}$ (mVμm)	$A_{0,\Delta\beta/\beta}$ (%μm)	$\rho(\Delta V_T, \Delta\beta/\beta)$ (-)
		NMOS	
no halo	3.08 ± 0.42	0.92 ± 0.10	0.15 ± 0.12
reference	4.01 ± 0.32	1.04 ± 0.08	0.30 ± 0.21
increased dose	5.53 ± 0.92	1.58 ± 0.39	0.62 ± 0.24
increased angle	5.20 ± 0.30	1.46 ± 0.23	0.51 ± 0.23
		PMOS	
no halo	2.25 ± 0.27	0.89 ± 0.11	0.02 ± 0.19
reference	2.51 ± 0.26	0.98 ± 0.11	0.13 ± 0.16
increased dose	3.13 ± 0.44	1.02 ± 0.16	0.27 ± 0.10
increased angle	2.97 ± 0.46	1.07 ± 0.13	0.27 ± 0.14

is present over the whole area of the transistor. We therefore believe that the gate does not act as a perfect mask and that part of the halos are implanted through the gate. This results in localized regions of high concentration of boron or arsenic at the gate side of the oxide or at the channel side, as is displayed in figure 5.9. Localized concentrations at the gate side result in extra fluctuations in gate depletion, while localized concentrations in the channel result in extra fluctuations in the

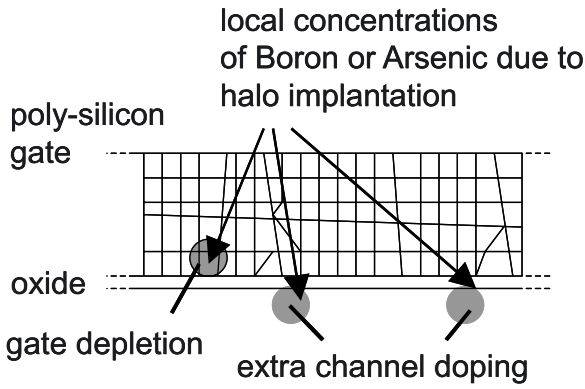


Figure 5.9. Schematic drawing of a MOSFET gate including localized regions of charge due to the halo implant. This charge can result in localized regions of extra gate depletion or in regions of extra charge in the channel.

threshold voltage and Coulomb scattering.
Now lets compare the different splits. As expected, it is found that the fluctuations increase when the halo dose is increased. Furthermore, it is found that the fluctuations become larger when the angle of the implant is increased. This can be due to two reasons. Firstly, it is possible that a 45° implantation results in charge that is located closer to the oxide-silicon interface than in the 35° case. It was found in section 4.3 that this would increase the fluctuations. Secondly, it is possible that 45° is a preferential direction with respect to channelling through a stack of grains or implantation along grain boundaries. Based on this, as future work it would be interesting to examine the increase in mismatch due to the halos as a function of the poly-grain size and structure of the gate.
Figure 5.8 shows the correlation between ΔV_T and $\Delta\beta/\beta$ as a function of the gate length. A positive correlation is observed, that increases with the amount of charge that is implanted through the gate. We will present two possible explanations for this positive correlation. For the first we need to take a look at figures 4.22 and 4.24, that display the calculated mismatch in ΔU_T and of its components as a function of the gate bias. The function ΔU_T was defined in such a way that its derivative gives the current factor, $\Delta\beta/\beta = -d\Delta U_T/dV_{GS}@V_{GS} - V_T = V_{ov}$, and that the intercept with the y-axis is equal to the mismatch in the threshold voltage, $\Delta V_T = \Delta U_T + (V_{GS} - V_T)\Delta\beta/\beta@V_{GS} - V_T = V_{ov}$. Looking at the figures, it is observed that the Coulomb scattering contribution to ΔU_T has a negative slope for overdrive voltages larger than $V_{ov} > 0.5$ V. This negative slope indeed causes a positive correlation

between ΔV_T and $\Delta\beta/\beta$. Note that applied gate-overdrive voltages generally lie around $V_{ov} \approx 0.3$ V. This indicates that our model for Coulomb scattering needs to be refined. Further note that we are not claiming that the mobility increases when more Coulomb scattering is present. We claim that Coulomb scattering is not properly accounted for in our strong-inversion drain-current model. As a result, it lowers the current by apparently increasing the threshold voltage, which is somewhat compensated for by an apparent increase in the current factor.
The second explanation for the positive correlation between ΔV_T and $\Delta\beta/\beta$ considers the possibility that our original explanation for the increase in the mismatch was not correct. The correlation could be an artifact of the fact that the drain current model on which the three-points extraction method is based was not derived for devices with halos. However, if this would be a problem, the correlation would be expected to decrease for longer transistors, which is not observed. In order to examine the exact influence of halos on device behavior, two dimensional simulations could be employed.
Finally, we note that, in parallel to this work, it was found in [120] that also LDDs can be implanted through the gate. It was shown that the matching performance of a 0.25 μm process can be improved by 1) reducing the LDD implantation energy, 2) increasing the thickness of the gate and 3) increasing the thickness of the implantation oxide on top of the gate. It was also reasoned that the matching performance can be improved by using poly-silicon as gate material instead of an amorphous gate.
In conclusion it can be stated that one has to be careful whenever the gate is assumed to act as a mask for an implantation step. When such an implantation goes through the gate, it can seriously degrade the matching performance of a technology.

5.3.2 Short- and narrow-channel effects

This subsection experimentally investigates the effect of the halo implantation on the short- and narrow-channel behavior of the matched parameters. Figure 5.10 shows $\sqrt{WL}\sigma_{\Delta V_T}$ and $\sqrt{WL}\sigma_{\Delta\beta/\beta}$ as a function of W/L for the experimental splits listed in table 5.4. All the measured devices have the same area of approximately $WL = 1.6$ μm^2. The shortest measured device is 0.13 μm long, while the narrowest device is 0.15 μm wide. In case of absence of short- and narrow-channel effects, the results plotted in figure 5.10 should yield horizontal lines. This is clearly not the case.
Comparing the different splits, it is observed that the mismatch increases for square devices when halos are implanted. This is because the halos

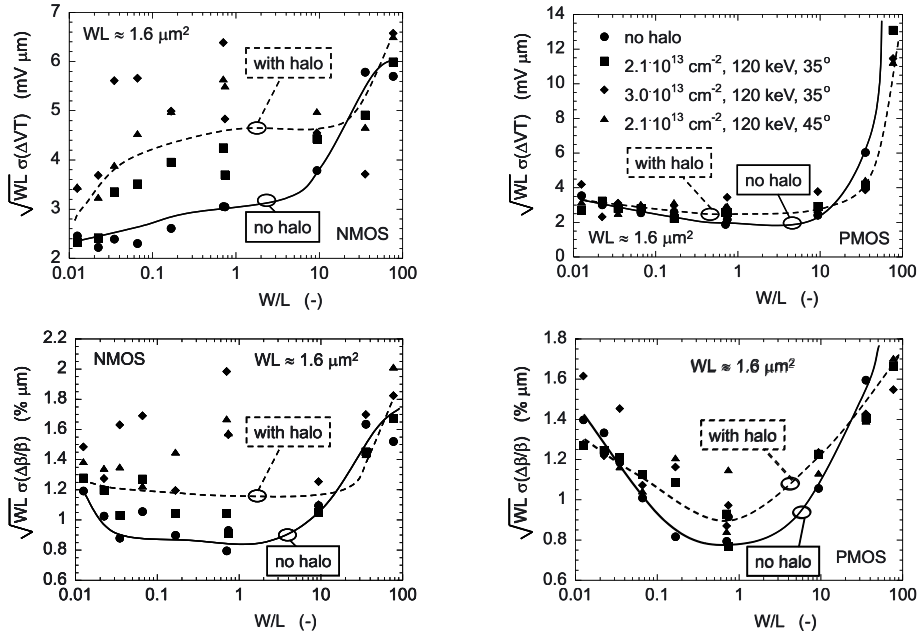


Figure 5.10. $\sqrt{WL}\sigma_{\Delta V_T}$ and $\sqrt{WL}\sigma_{\Delta\beta/\beta}$ as a function of W/L for the experimental splits listed in table 5.4. Symbols have the same meaning as in figure 5.7. The lines are introduced to guide the eye.

are implanted through the gate, as was reasoned in the previous subsection. We will now consider the deviations for the short transistor pairs, thus the pairs with high W/L-ratio. As was mentioned in subsection 4.2.3, deviations from the $\sigma_{\Delta P} \propto 1/\sqrt{WL}$ relationship can originate from four effects, namely: 1) A shorter effective channel length than the metallurgical length, 2) the increase in surface potential, caused by the proximity of the extension regions, 3) the increase in doping level due to the halos, and 4) fluctuations in the short-channel effects themselves. Looking at figure 5.10, it is observed that the relative increase in mismatch for short-channel transistors is more prominent when no halos are implanted, because devices with halos possess a larger effective channel length. In case of the NMOS transistors without halos, it is observed that for the shortest gate length the matching performance becomes slightly better again. If significant, this could be explained as follows: When the gate length becomes very short, the main difference between the splits will no longer be determined by the effective channel length, but by the amount that the potential barrier between source and drain is lowered. This barrier lowering is strongest for the transistors with-

out halos. This behavior is not observed for the PMOS transistors. Note that the PMOS transistors suffer more from short-channel effects than the NMOS transistors in this technology and fluctuations in the short-channel effect itself might start to play a prominent role. These fluctuations are expected to be strongest for the transistors without halos.

We will now consider the narrow-channel effects, displayed in figure 5.10 by the devices with low W/L-ratio. As was seen in subsection 4.2.3, the threshold voltage mismatch for the NMOS transistors decreases with decreasing width, while a slight increase is observed for the PMOS transistors. The mismatch in the current factor increases for both NMOS and PMOS transistors. Furthermore, it is observed that the difference in mismatch between devices with and without halo, decreases as the width is reduced. A possible explanation is that the STI introduces topography in the poly-silicon gate, which effectively reduces the amount of halo charge that is implanted through the gate.

5.4 Comparison of different CMOS technologies

By comparing the matching properties of several CMOS processes, it was found in [6] that, as a rule of thumb, $A_{0,\Delta V_T}$ in mV·μm is equal to the effective oxide thickness in nm for a well optimized process. Current-factor mismatch was found to be independent of the technology generation and $A_{0,\Delta\beta/\beta} = 1.0$ %μm.

In order to try to understand these empirical laws we first need to assume some scaling relationships. Lets assume that $t_{ox,eff} \propto \kappa^{-1}$, where κ is the scaling factor. In this case the doping concentration $N_A \propto \kappa^{1.5}$, the depletion-layer width $W_D \propto \kappa^{-0.75}$ and the inversion-layer width $z_\mu \propto \kappa^{-0.5}$. The inversion-layer charge at constant gate overdrive scales as $Q_i \propto \kappa$. These are crude approximations, but they will serve for the purpose of an order of magnitude calculation.

In figure 4.22a out of subsection 4.3.6 it was observed that the mismatch in the drain current[2] at low gate overdrive is mainly determined by fluctuations in channel doping. The fluctuation in channel doping affects the drain current through threshold-voltage fluctuations and fluctuations in the amount of Coulomb scattering, as was observed in figure 4.24a. From (4.79), it directly follows that fluctuations in the threshold voltage scale like $A_{0,\Delta V_T} \propto \kappa^{-0.625}$. To estimate the scaling of the fluctuations in Coulomb scattering, it is assumed that $\mu/\mu_C \propto z_\mu N_A/Q_i$. Together with (4.101) this yields $A_{0,\Delta U_T}|_{\Delta\mu_C} \propto \kappa^{-0.5}$. We conclude that the two

[2]Remember that $\sigma_{\Delta I_D/I_D} = \sigma_{\Delta U_T}/(V_{GS} - V_T)$.

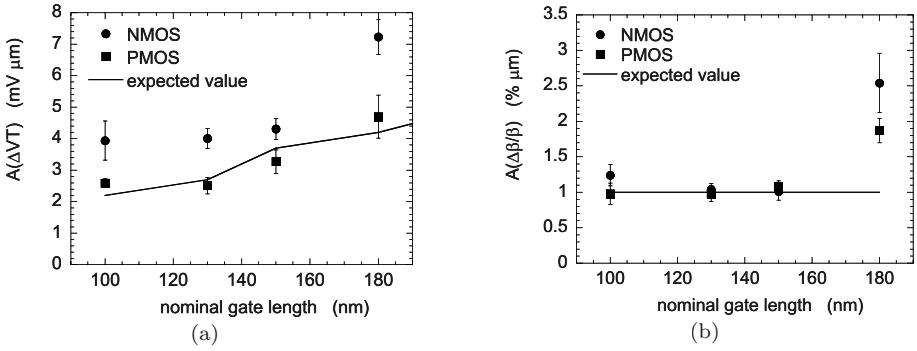


Figure 5.11. $A_{0,\Delta V_T}$ (a) and $A_{0,\Delta\beta/\beta}$ (b) for several technology generations as a function of their nominal gate length. Error bars are 1σ long. The solid lines plot the expected values: $A_{\Delta V_T}$ in mVμm is expected to be equal to the effective oxide thickness in nm. The physical oxide thicknesses of the investigated technologies are listed in table 5.1.

main components to the mismatch scale less strongly than the oxide thickness, opposing the experimental observations. This suggests that people learn from mistakes in older technologies, which helps to improve the newer ones. It also indicates that research remains necessary to keep on obtaining the same levels of improvement with down scaling.

Now consider the current factor. It is observed in figures 4.22b and 4.24b+d that the current factor is determined by several relatively small mechanisms. Some of them decrease when the transistor is scaled down, while others increase. From this, it can be understood that the overall mismatch in the current factor did not drastically differ for different technology generations. Lets consider the fluctuations in the gate doping separately, since they contribute the most. It follows from (4.85) and (4.87) that approximately $\sigma_{\Delta\beta/\beta}|_{\Delta N_p} \propto (Q_D + Q_i)\sigma_{\Delta N_p/N_p}/t_{ox,eff}N_p$. When we assume that the doping concentration in the gate and $A_{0,\Delta N_p/N_p}$ do not scale, it follows that $A_{0,\Delta\beta/\beta}|_{\Delta N_p}$ scales at a rate of $\kappa^{1.75} - \kappa^{2.0}$. This clearly demonstrates the need for the scaling of $A_{0,\Delta N_p/N_p}$, which is generally done by reducing the poly-grain size, and ultimately by introducing metal gates.

We will now look further into how well the technologies discussed in this chapter follow the empirical scaling laws. Note that none of these technologies was optimized with respect to their matching performance. Figure 5.11 shows $A_{0,\Delta V_T}$ and $A_{0,\Delta\beta/\beta}$ as a function the nominal gate length of the technologies presented in table 5.1. Reference processing conditions were used. The solid line shows the expected value. It is observed that the technology with $L_{nominal} = 180$ nm has a significantly

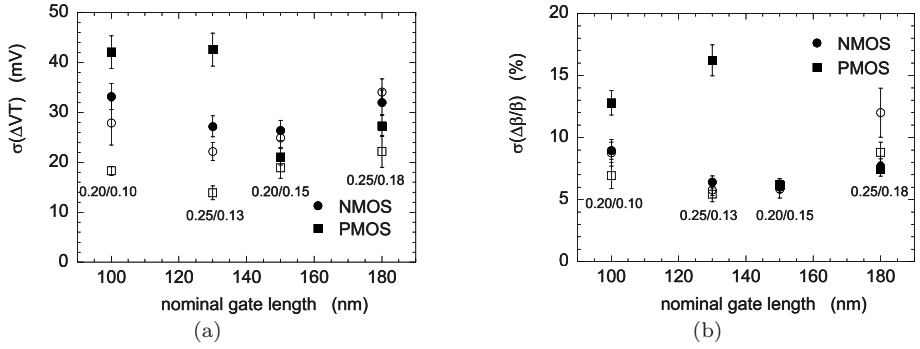
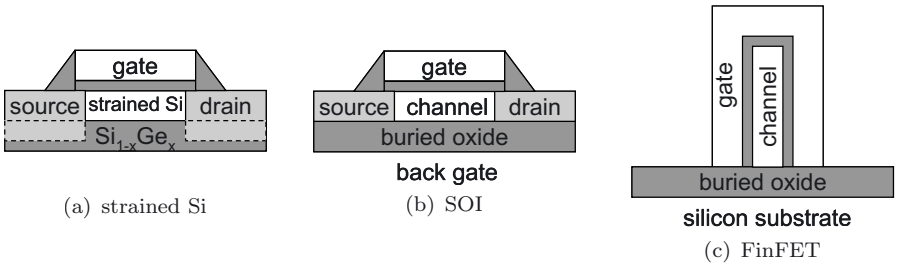


Figure 5.12. $\sigma_{\Delta V_T}$ (a) and $\sigma_{\Delta\beta/\beta}$ (b) for the minimum size transistor of several technology generations as a function of their nominal gate length. The W/L-ratios are included in the figure and are in (μm/μm). Solid symbols are based on experimental data, while the values of the open symbols are calculated from $A_{0,\Delta V_T}$ (a) and $A_{0,\Delta\beta/\beta}$ (b) and the device areas. Error bars are 1σ long. The physical oxide thicknesses of the investigated technologies are listed in table 5.1.

worse matching performance than later technologies. This is caused by the change in gate material from amorphous- to poly-silicon. In case of $A_{0,\Delta V_T}$, the PMOS transistors perform as expected. For the NMOS transistors, the observed matching properties lie above the expected values. This could be due to halos that are unintentionally implanted through the gate. Note that the poly-grain size was not scaled. In case of $A_{0,\Delta\beta/\beta}$, it is observed that the mismatch indeed lies around 1.0 %μm for technologies with poly-silicon gates.

In [8] it is calculated that, in order to obtain a 90 % yield on a 1 Mbit SRAM, it is required that $A_{0,\Delta V_T} < 6.0$ mVμm for an $L_{nominal} = 180$ nm technology and that $A_{0,\Delta V_T} < 2.5$ mVμm for an $L_{nominal} = 100$ nm technology. It is observed in figure 5.11a that for the NMOS transistors these specs are not reached. The picture becomes worse when we look at the matching properties of the minimum size transistor, which is displayed in figure 5.12 for the examined technology generations. In this figure the experimental data is represented by solid symbols. The open symbols are calculations based on the area of the minimum transistor and on $A_{0,\Delta V_T}$ and $A_{0,\Delta\beta/\beta}$. The mismatch in PMOS transistors is seriously increased because of short-channel effects. In order to improve SRAM yield, these short-channel effects need to be brought under control. The mismatch of the NMOS transistors might also be improved by reducing the poly-grain size. Finally note that one can also decrease the sensitivity of SRAM yield to parameter fluctuations by increasing the threshold voltage of the SRAM transistors by an extra implantation step. How-



(a) strained Si (b) SOI (c) FinFET

Figure 5.13. Schematic representation of a MOSFET with a strained silicon channel (a), a silicon-on-insulator (SOI) MOSFET (b) and a FinFET (c). The source and drain of the FinFET are located in front and to the back of the channel, respectively.

ever, this solution requires an extra mask and lithography step, which increases the cost of the process.

5.5 Alternative device concepts

Ever since the semiconductor industry started with the down-scaling of device dimensions, people have tried to predict the end of this down-scaling. These predictions are based on economical, technological and physical considerations and the relationships between them. However, it is wrong to assume that everything will end at a fixed point. It is better to talk about a gradual decline or, to put it more positively, change. For example, because of economical reasons, a lot of companies stopped to develop new processes or they even went fabless. However, semiconductor foundries took advantage of this situation and started to grow more strongly. A constant technological issue is the development of new lithography tools with rising costs. Physical barriers started to play a role when the supply voltage needed to be scaled down, because of the reliability of the gate oxide.

At this moment one of the main concerns and physical barriers is caused by the leakage current, which can result in unacceptably large power dissipation. To keep leakage levels under control, threshold-voltage scaling slowed down. Together with a decreasing supply voltage, this would result in a lowering of the drive current. To solve this problem, alternative device concepts are introduced. For example, one can use germanium to strain the silicon (see figure 5.13a), which increases the mobility. Another option is to use a silicon-on-insulator (SOI) substrate instead of a standard silicon one (see figure 5.13b). This reduces the junction capacitances, which increases the speed. Another possibility is the use of FinFETs, which are double- or triple-gate devices (see figure 5.13c). Because of the multiple gates, the drain current increases. An-

other advantage of SOI devices and FinFETs is that they allow for a better control of short-channel effects.
The question to ask ourselves is how this change in device structure affects the matching behavior. As an introduction, some of the issues will be briefly discussed.

Strained silicon MOSFETs. A strained silicon MOSFET contains a silicon-germanium layer on which a thin silicon layer ($t_{Si} \sim 10$ nm) is grown epitaxially. The silicon is strained, because it has a higher lattice constant than the relaxed $Si_{1-x}Ge_x$ layer, where x is the fraction of Germanium atoms. Besides the mobility, the difference in lattice constants causes a band offset, which results in a negative shift of the threshold voltage in the range of $0-500$ mV. This shift is a function of the germanium concentration and the thickness of the strained silicon film, which could give rise to extra parameter fluctuations [121]. The device is most sensitive to these fluctuations when the depletion layer width is equal to the silicon film thickness.
To compensate the negative shift in threshold voltage, for the NMOS transistors extra doping needs to be implanted. This results in an increase in the parameter fluctuations. Furthermore, because of the increase in bulk mobility, the contribution of other mobility components becomes more dominant, and therefore also their contribution to the amount of fluctuations. Fluctuations in the bulk mobility itself are also expected to increase, because of fluctuations in the germanium content and film thickness. Overall, we conclude that a strained silicon MOSFET will be more difficult to optimize with respect to parameter fluctuations than a standard MOSFET.

Silicon-on-insulator MOSFETs. The channel region of an SOI MOSFET consists of a thin silicon layer on top of an oxide. When this silicon layer is thicker than the depletion-layer width, the SOI MOSFET is called partially depleted. The fabrication and operation of such a MOSFET is very similar to that of a normal MOSFET and no major changes in the matching behavior are expected. When the silicon layer is thinner than the depletion-layer width, the SOI MOSFET is called fully depleted (FD). In this case a higher doping concentration is needed to reach the required threshold voltage. In a fully depleted device an increase in the doping level is not compensated for by a decrease in depletion layer width, which makes the device more sensitive to doping fluctuations. The threshold voltage is also dependent on the silicon-film thickness, which could cause an increase in the fluctuations.
As was mentioned before, using a FD-SOI device improves short channel behavior. Actually, when the silicon film is thin enough, no doping would be required at all. In this case the threshold voltage is determined

by the gate work-function, which needs to be properly engineered. This is not straightforward, but it solves the problem of doping fluctuations. The sensitivity of such a device to film thickness and gate length was examined in [122]. It was also found that these sensitivities decrease when a negative voltage is applied to the back gate.

FinFETs. FinFETs are fabricated by etching silicon fins on an SOI substrate. The gate goes around the whole fin, making it a double- or triple-gate device. The device is contacted at the front and at the back (not shown in the two dimensional figure 5.13c). The sensitivity of FinFETs to device fluctuations was examined in [122, 123]. The same issues play a role as for FD-SOI MOSFETs. However, in the case of FinFETs the silicon film thickness is determined by the lithography and etching processes, which means that it is less well controlled than the silicon film thickness of SOI devices. Also, the gate oxide is grown on the side of the fins, which will result in a worse control of its thickness.

5.6 Conclusions

This chapter presented some examples of how certain process steps can influence the matching behavior of a technology. Furthermore, the scaling of matching properties was discussed and issues for alternative device structures were briefly addressed.

It was found that decreasing the grain size of the poly- or amorphous-silicon gate material can greatly improve the matching behavior of a technology, because it reduces the correlation length of the mismatch causing stochastic process. By comparing devices with halo implantation and without halo implantation, it was found that the halo seriously degrades the matching performance, which was mainly observed for the NMOS transistors. The gate did not act as a perfect mask for the implantation step and the halo was implanted through the gate. This resulted in extra localized charge in the channel region or at the oxide-gate interface. Increasing the halo dose or the implantation angle worsens the effect. However, for short devices, the halo improves the matching behavior, because of the increase in effective channel-length. For very short NMOS transistors, the devices without halo start to perform better again. For a very short device the effective channel length is no longer determined by the doping concentration, but the higher doping concentration in the devices with halos causes a decrease in the matching performance. For narrow devices the impact of the halo becomes smaller. This could be due to the the topography introduced by the shallow-trench isolation, that can scatter the implantation.

With respect to the scaling of the matching properties of CMOS technologies, it is concluded that the matching performance improves faster

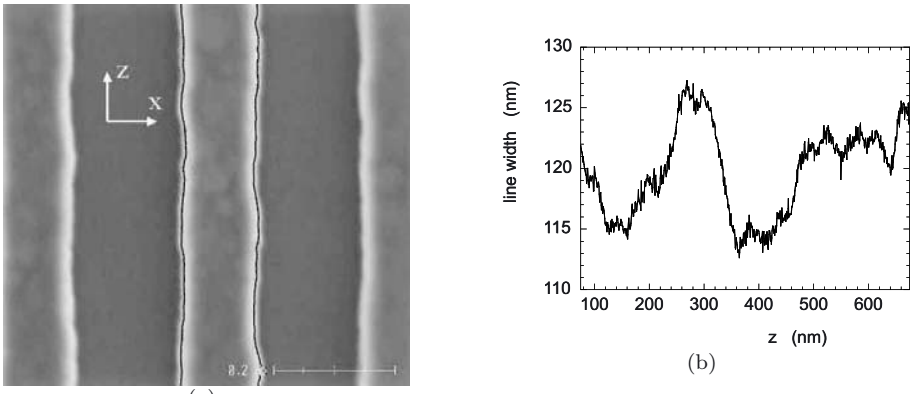
with scaling than what is expected theoretically. This indicates that newer technologies profit from what is learned during the development of older technologies. However, it also indicates that research efforts remain necessary to keep parameter fluctuations under control.
The matching performances of four experimental technologies are compared to what is expected based on literature. The nominal gate lengths of these technologies ranged from 100 nm to 180 nm. It was found that the matching performance significantly improved after changing from amorphous to poly-silicon gate material. Furthermore, the PMOS transistors were observed to follow the empirical scaling laws. The NMOS transistors have worse matching performance than empirically expected. This could be improved by reducing the grain size of the gate material or by making sure that the halos are not implanted through the gate. For the minimum size transistors, the matching performance of the PMOS devices is also poor. This is caused by the poor short-channel control of these transistors, which needs to be improved in order to be able to obtain acceptable SRAM yield.

Chapter 6

IMPACT OF LINE-EDGE ROUGHNESS ON PARAMETER FLUCTUATIONS, OFF-STATE CURRENT AND YIELD

Doping fluctuations are considered to determine one of the fundamental lower limits to parameter fluctuations. These effects have been extensively studied in literature and in chapter 4. A less well studied effect is the impact of line-edge roughness (LER). While the previous chapter discussed some current technology issues, LER is considered to be one of the main limiting factors for future technologies. In general, the printed gates of transistors exhibit a certain roughness. This roughness is ultimately limited by Poisson statistics on the number of photons during the exposure of the resist [124]. However, in practice, chemical properties of the resist make out the main contribution to LER [125]. As transistor gate-lengths are scaled down, LER is expected to have an impact on *parameter fluctuations*, *off-state current* and *yield*.
In literature, the effects of LER have mainly been investigated by 2D [126, 127] or 3D [128, 129] device simulations. In the case of 2D simulations, the poly-gate is divided in small segments and for each segment the current is found from the simulation. The same approach is followed in [130], but here an analytical model is used to describe the drain current of a segment. In [127][1] the simulations were calibrated on a 0.13 μm process on which the LER was exaggerated. These simulations were then used to fix the requirements on LER for 34 nm gate-length transistors. However, it is not investigated how the importance of LER increases as MOSFETs scale down and when it will really become an issue. This is essential information for gate-patterning process optimization as well as device optimization.

[1]The work presented in this publication was done in parallel to the work presented in this chapter.



(a) (b)

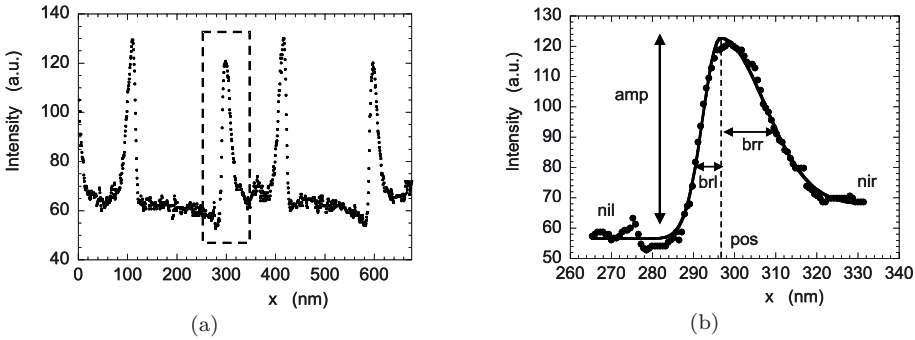
Figure 6.1. a) SEM picture of three poly-silicon lines after etch. The edges of the fully displayed center line have been highlighted. b) Local line width as a function of the position along the line (z).

In this chapter we will start in section 6.1 by characterizing the LER itself. In section 6.2, the theory presented in section 2.4 and section 4.2 will be used to describe the impact of LER on parameter fluctuations, off-state current and yield. This will result in analytical expressions that directly link properties of the LER to the above mentioned effects. In section 6.3 these models are experimentally verified, which, in section 6.4, allows us to make predictions of the impact of LER and to present guidelines for LER scaling. Section 6.5 concludes this chapter.

6.1 Characterization of line-edge roughness

When examining mismatch effects, generally one does not have any directly measurable information about the stochastic microscopic processes that cause the mismatch. Line-edge roughness forms an exception, because the lines are visible after gate patterning, and stochastic properties can be extracted. Figure 6.1a shows a SEM picture of a printed poly-silicon line for which the two edges are highlighted. Figure 6.1b plots the local line width as a function of the position along the line (z).

To detect the edge, in figure 6.2a we look at the intensity profile of the SEM picture at a certain position along the line. Four peaks are observed, that are related to the four line-edges in the picture. Figure 6.2b zooms in on one of these peaks, which is fitted by a function that



(a) (b)

Figure 6.2. a) Intensity profile of the SEM picture displayed in figure 6.1a at a certain value of z b) Zoom-in on one of the peaks.

consists of two halves of Gaussian functions:

$$Intensity = \begin{cases} nil + amp \cdot e^{-((x-pos)/brl)^2} & x < pos \\ nir + (amp - nir + nil) \cdot e^{-((x-pos)/brr)^2} & x > pos \end{cases} . \tag{6.1}$$

The fitted parameters *amp*, *brl*, *brr*, *nil*, *nir* and *pos* are defined in figure 6.2b. The edge of the line is assumed to be located at the point of maximum intensity, thus at $x = pos$. According to the ITRS roadmap [109], the standard deviation of this edge position (σ_{LER}) should be smaller than 3.3 % of the gate length, but no justification is presented for this number. Furthermore, figure 6.1 contains information about the shape of the line, which is neglected when only σ_{LER} is taken into account, but which is required to describe the impact of LER on transistor behavior. Also the correlation between the roughness of the two edges can be calculated. However, it is found to be insignificant ($\rho = 0.19 \pm 0.32$). This means that the variance of the line-width roughness (LWR) is equal to two times the variance in LER, i.e. $\sigma^2_{LWR} = 2\sigma^2_{LER}$. From this point on, we choose to analyze LWR instead of LER. This has the advantage that slight rotations in an analyzed SEM picture are to first order cancelled out.

The full spectral properties of the LWR are contained in its autocovariance function (R_{LWR}), which is the Fourier transform of its power spectrum, and is defined by:

$$R_{LWR}(d) \equiv \sigma^2_{LWR} \cdot \rho(L_{local}(z), L_{local}(z+d)), \tag{6.2}$$

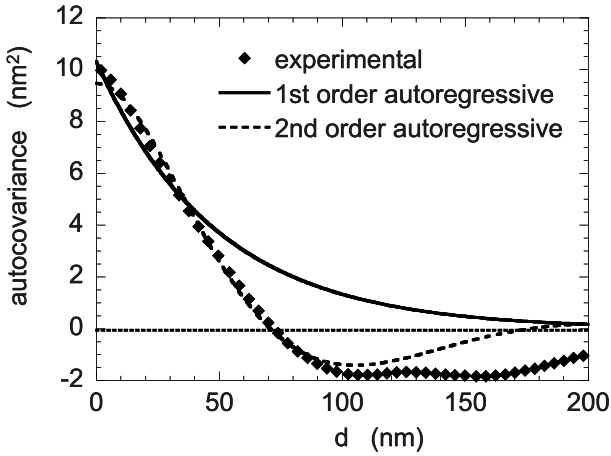


Figure 6.3. Autocovariance function of the LWR of a 193 nm gate-paterning process. A first-order and second-order autoregressive model are fitted to the experimental data.

where the autocorrelation function $\rho(L_{local}(z), L_{local}(z+d))$ is the correlation of the width of the line[2] at a certain position z and its width at a certain position $z+d$ further down the line. Note that $R_{LWR}(0) = \sigma^2_{LWR}$ and that the width of this function presents a measure of the correlation width of the process.
The autocovariance function is calculated from experimental data by the biased estimate as follows:

$$R_{LWR}(n \cdot step) = \frac{1}{N} \sum_{i=1}^{N-n} (L_{local}[i] - \overline{L_{local}})(L_{local}[i+n] - \overline{L_{local}}). \quad (6.3)$$

These experimental data consist of N measurements of the local length at a distance *step* from each other. Figure 6.3 shows the autocovariance function of a state-of-the-art 193 nm lithography process. This autocovariance function is extracted from 5 lines with a length of 700 nm each. Also, fits are shown of a first-order and second-order autoregressive model. The autocovariance function of a first-order autoregressive process is given by:

$$R_{LWR}(d) = \sigma^2_{LWR} \cdot e^{-\alpha_1 |d|}. \quad (6.4)$$

For the examined lithography process it is found that $\sigma_{LWR} = 3.2$ nm and $\alpha_1 = 0.020$ nm^{-1}. The autocovariance function of a second-order

[2]The local width of the line is equal to the local length of the gate.

autoregressive process is given by:

$$R_{LWR}(d) = \sigma_{LWR}^2 \cdot e^{-\alpha_2|d|}\left(\cos(p|d|) + \frac{\alpha_2}{p}\sin(p|d|)\right). \tag{6.5}$$

For the lithography process under study, the fit yields $\sigma_{LWR} = 3.1$ nm, $\alpha_2 = 0.036$ nm^{-1} and $p = 0.030$ nm^{-1}. It is observed that the second-order process gives the best fit. However, since the data plotted in figure 6.3 is only accurate at the lowest values of d, we cannot justify using the second-order process. Furthermore, a second-order autoregressive process indicates some kind of damped oscillation in space related to the LER causing process, which is not physical. We will find in the next section that parameter fluctuations due to LER are proportional to the total area under the autocovariance function, which is larger for the first-order process. Since we want to use our calculations to provide upper boundaries for the maximum allowable LER, we choose to use the first-order autoregressive process to describe the roughness of the line.

6.2 Modeling the impact of line-width roughness

Most transistor parameters are a function of its length, as is displayed in figure 6.4 for the threshold voltage and the off-state current[3,4]. Line-width roughness influences transistor behavior through this length dependence. In order to calculate its impact, we will use the one-dimensional equivalent of the theory published in [5], which was summarized in section 2.4. As example of a strong-inversion parameter, the impact of LWR on the threshold voltage mismatch is calculated in subsection 6.2.1[5]. It was found in subsection 4.2.1 that deviations are to be expected in the weak-inversion regime and the impact of LWR on the off-state current is calculated in subsection 6.2.2. Subsection 6.2.3 models the impact of LWR on yield.

[3]The off-state current can also be referred to as the leakage current.
[4]All experimental results shown in this section are for a 130 nm technology with a nominal gate length of 100 nm, an oxide thickness of 1.5 nm and a supply voltage of 1.2 V.
[5]For other strong inversion parameters, like e.g. the on-state current, the same approach can be followed.

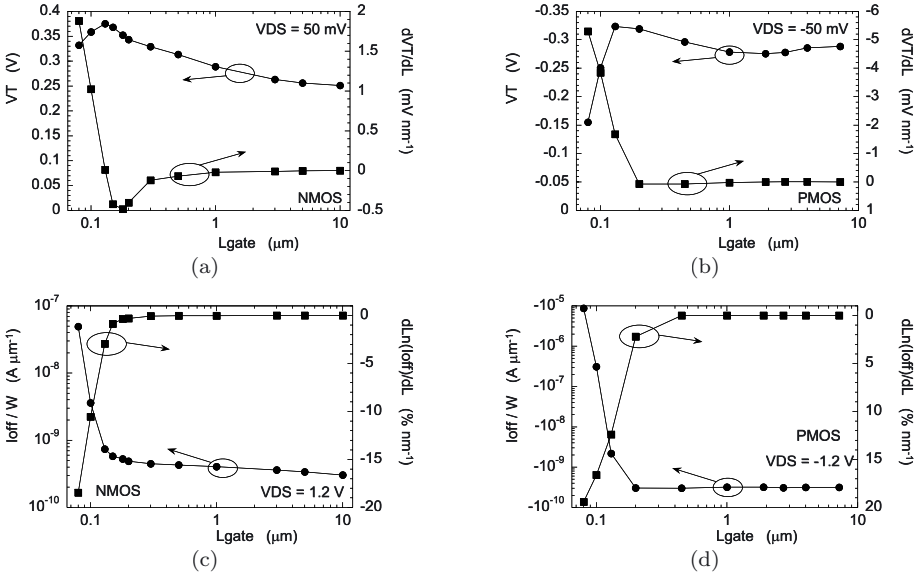


Figure 6.4. left axes: Threshold voltage (a+b) and off-state current (c+d) as a function of the gate length. Right axes: the derivative of the threshold voltage (a+b) and of the logarithm of the off-state current (c+d) to the gate length. The oxide thickness is equal to 1.5 nm.

6.2.1 Impact of line-width roughness on the threshold voltage

Locally, i.e. at a certain value of z, the variation in threshold voltage due to LWR is calculated by:

$$\sigma_{\Delta V_{T,local}} = \sqrt{2} \left| \frac{dV_T}{dL} \right| \sigma_{LWR}. \tag{6.6}$$

The overall variation in the mismatch of the threshold voltage is calculated by averaging the local variations over the width of the transistor. Mathematically this translates into:

$$\sigma^2_{\Delta V_T} = 2 \left(\frac{dV_T}{dL} \right)^2 [G * G * R_{LWR}](0) = \tag{6.7}$$

$$= 2 \left(\frac{dV_T}{dL} \right)^2 \frac{2}{\alpha_1 W} \left(1 - \frac{1}{\alpha_1 W} \left(1 - e^{-\alpha_1 W} \right) \right) \cdot \sigma^2_{LWR},$$

where the geometry function $G(z) = 1/W$ for $|z| < W/2$ and $G(z) = 0$ for $|z| > W/2$. The last equality holds when the LWR is described by a

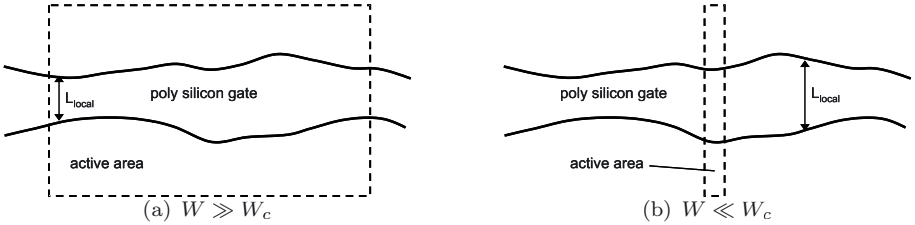


Figure 6.5. Schematic top-view of a transistor with a much larger width than the correlation width of the LWR (a) and of a transistor with a much smaller width than this correlation width (b).

first-order autoregressive model. Two extreme cases are distinguished. Firstly, when the transistor is much wider than the correlation width (W_c) of the LWR, as is schematically displayed in figure 6.5a, then:

$$\sigma^2_{\Delta V_T} \cong \frac{2}{W}\left(\frac{dV_T}{dL}\right)^2 \int_{-\infty}^{\infty} R_{LWR}(z)dz = \frac{W_c \cdot \sigma^2_{\Delta V_{T,local}}}{W}, \tag{6.8}$$

from which it follows that W_c is equal to the area under the autocorrelation function[6]. Since the maximum of the autocorrelation function is equal to 1, the total area under this function is a measure of its width. When the LWR is represented by a first order autoregressive process, then $W_c = 2/\alpha_1$, which is equal to $W_c = 100$ nm for the case displayed in figure 6.3.
Secondly, in the other extreme case when $W \ll W_c$, as is displayed in figure 6.5b, (6.7) can be approximated by:

$$\sigma^2_{\Delta V_T} \cong 2\left(\frac{dV_T}{dL}\right)^2 R_{LWR}(0) = \sigma^2_{\Delta V_{T,local}}. \tag{6.9}$$

This means that the mismatch causing process does not have enough space to change the local length over the width of the transistor and the variation between transistors is equal to the local variation.
Figure 6.6 compares the experimentally obtained mismatch in threshold voltage with the calculated mismatch due to LWR. The $\sigma_{\Delta V_T}$ is normalized to $1/\sqrt{W}$. It is observed that LWR does not give a significant contribution to the fluctuations down to the minimum available gate length of 80 nm.

[6]Note that the total area under the autocorrelation function is equal to the zero-frequency component of the normalized power spectrum.

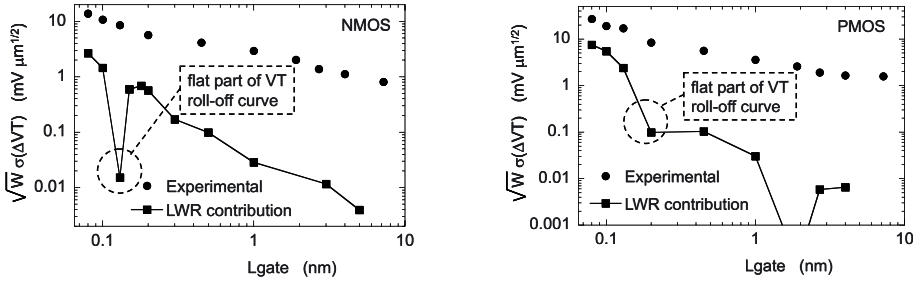


Figure 6.6. Experimental mismatch in the threshold voltage (symbols) and calculated mismatch due to LWR (lines). The calculations use the first order autoregressive model to describe the LWR for which the parameters are extracted from figure 6.3.

6.2.2 Impact of line-width roughness on the off-state current

The off-state current varies much more strongly with the gate length than the threshold voltage, as can be seen in figure 6.4. For this reason a linearization like in (6.6) is not accurate. However, the logarithm of the off-state current can be linearized:

$$\sigma_{ln} \equiv \sigma_{\Delta' ln(I_{off}),local} = \left| \frac{dln(I_{off})}{dL} \right| \sigma_{LWR}. \tag{6.10}$$

Locally, this results in a lognormal distribution of the off-state current due to LWR. In our analysis we consider the ratio of the off-state current of a device that suffers from LWR and the off-state current of an ideal device with no LWR. This is denoted as $r'I_{off}$.
Firstly, consider the average of this ratio ($\mu_{r'I_{off}}$), which ideally is equal to one. However, the asymmetry of the lognormal distribution causes an increase in the off-state current, which is given by:

$$\mu_{r'I_{off}} = e^{\sigma_{ln}^2/2}, \tag{6.11}$$

and is independent of the transistor width. Using $\sigma_{LWR} = 3.2$ nm and figure 6.4, it follows that an increase of about 20 % is expected for the $L_{gate} = 80$ nm transistors, which is relatively small.
Secondly, consider the fluctuations in $r'I_{off}$. Because of the nonlinearity introduced by the lognormal distribution, the shape of the autocovariance function of this ratio ($R_{r'I_{off}}(d)$) is not the same as that of the LWR. In order to calculate $R_{r'I_{off}}(d)$ we will assume that the current only flows in the x-direction. Then, by using (3.4), it is found that:

$$R_{r'I_{off}}(d) = e^{\sigma_{ln}^2} \cdot \left(e^{\rho_{LWR}(d) \cdot \sigma_{ln}^2} - 1 \right), \tag{6.12}$$

where $\rho_{LWR}(d)$ is the autocorrelation function of the LWR. The variance in $r'I_{off}$ is now given by:

$$\sigma^2_{r'I_{off}} = [G * G * R_{r'I_{off}}](0). \tag{6.13}$$

For wide transistors ($W \gg W_{c,I_{off}}$) this equation simplifies to:

$$\sigma^2_{r'I_{off}} \cong e^{\sigma^2_{ln}} \cdot \left(e^{\sigma^2_{ln}} - 1\right) \frac{W_{c,I_{off}}}{W}, \tag{6.14}$$

where $W_{c,I_{off}} \equiv \int_{-\infty}^{\infty} R_{r'I_{off}}(z)/R_{r'I_{off}}(0)dz$. For the first order autoregressive process shown in figure 6.2, $W_{c,I_{off}} = 90 - 100$ nm, depending on the sensitivity of the off-state current to the gate length as displayed in figure 6.4.
For narrow transistors (6.13) simplifies to:

$$\sigma^2_{r'I_{off}} \cong e^{\sigma^2_{ln}} \cdot \left(e^{\sigma^2_{ln}} - 1\right). \tag{6.15}$$

It is not straightforward to compare this theory to experiment, because $r'I_{off}$ is defined with respect to an ideal device of which we do not know the electrical properties. This problem is overcome by matching two transistors in such a way that the ideal device drops out of the equation. This is obtained by analyzing the mismatch in the logarithm of the off-state current ($\Delta ln(I_{off})$). In order to calculate the variance of this quantity from (6.13), the distribution of $r'I_{off}$ needs to be known. For a very narrow device this distribution is expected to be lognormal, while it is normal for a very wide device, as follows from the central limit theorem. However, generally the distribution lies somewhere in between these extreme cases. As an approximation, we will assume $r'I_{off}$ to be lognormally distributed with a mean given by (6.11) and a variance given by (6.13). This results in:

$$\sigma^2_{\Delta ln(I_{off})} \approx 2ln\left(\sigma^2_{r'I_{off}} \cdot e^{-\sigma^2_{ln}} + 1\right). \tag{6.16}$$

This equation gives correct results for very narrow and very wide transistors. For the intermediate cases the inaccuracy has been checked in a numerical way and it is found to be smaller than 20 % when $\sigma^2_{ln} < 1.0$. Figure 6.7 compares the experimentally obtained mismatch in the off-state current with the calculated mismatch due to LWR. Like for the threshold voltage, it is observed that LWR does not give a significant contribution to the fluctuations down to the minimum available gate length of 80 nm.

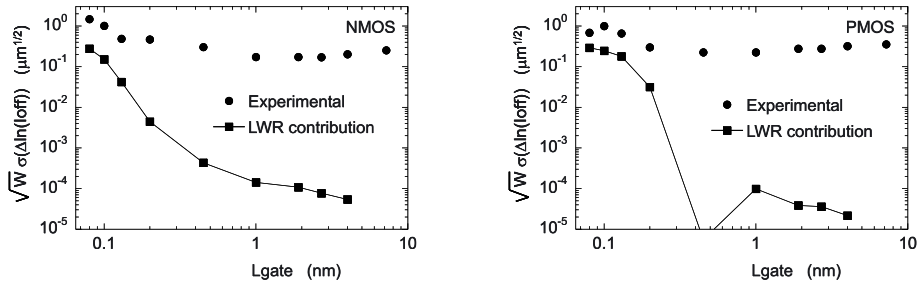


Figure 6.7. Experimental mismatch in the logarithm of the off-state current (symbols) and calculated mismatch due to LWR (lines). The calculations use the first order autoregressive model to describe the LWR for which the parameters are extracted from figure 6.3.

6.2.3 Impact of line-width roughness on yield

Besides causing parameter fluctuations and increasing the off-state current, LWR can also decrease yield. To calculate this decrease, it is assumed that a device fails when it is locally shorter than a certain critical gate length ($L_{critical}$). The probability (p_{local}) that this happens at a specific location follows from the normal distribution of the LWR and is given by:

$$p_{local} = \frac{1}{2} erfc\left(\frac{L_{gate} - L_{crit}}{\sqrt{2}\sigma_{LWR}}\right), \tag{6.17}$$

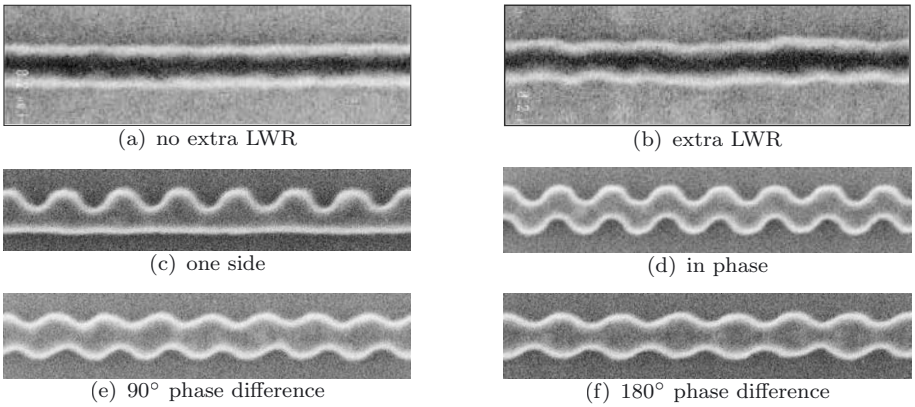
where $erfc$ is the complementary error function. We now assume that the device consists of W/W_c segments when $W > W_c$ and of one segment when $W < W_c$. Within one segment the line-width is constant, the standard deviation of the line width is equal to σ_{LWR}, and the deviation of one segment from the average is uncorrelated with the deviations of the other segments. From this it follows that the probability that one device fails (p_{device}) is equal to:

$$p_{device} = 1 - (1 - p_{local})^{min(1, W/W_c)}. \tag{6.18}$$

When a circuit contains N_{device} devices and the circuit fails when one transistor fails, the circuit yield is given by:

$$yield = (1 - p_{device})^{N_{device}}. \tag{6.19}$$

As an example, consider a 1Mbit SRAM, which has 6 million minimum-size transistors. We assume $W < W_c$ and $L_{crit} = 0.7 L_{gate}$. If we allow for a maximum yield loss due to LWR of 0.5 %, then the requirement



(a) no extra LWR (b) extra LWR

(c) one side (d) in phase

(e) 90° phase difference (f) 180° phase difference

Figure 6.8. Top-view SEM pictures of the gates of the special transistors that were fabricated using e-beam patterning to study LWR effects

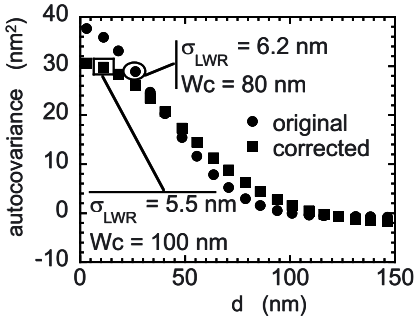
on LWR is $6\sigma_{LWR} < 0.3L_{gate}$, which is equal to the ITRS roadmap requirement. Requirements on LWR based on parameter fluctuations and the off-state current will be presented in section 6.4.

6.3 Experimental investigation of the impact of line-width roughness

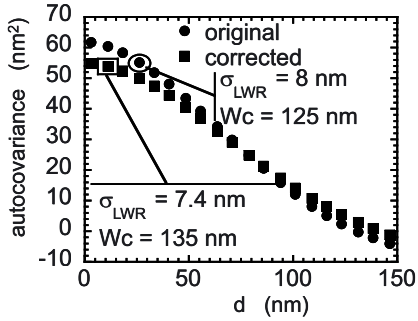
As was shown in the previous section, LWR does not significantly affect MOSFET behavior down to gate lengths of 80 nm. Therefore, in order to experimentally investigate LWR, it needs to be artificially increased. This will allow us to make predictions for future technologies. The setup of our experiments is described in the first subsection of this section. Results are given in the second and third subsections while the last subsection deals with the issue of yield.

6.3.1 Experimental setup

To create transistors with extra rough gates, electron-beam (e-beam) lithography was used. E-beam lithography has the advantage that it can produce any gate shape with a resolution of approximately 20 nm. This allowed us to create transistors with sinusoidally shaped gate-edges, as displayed in figures 6.8c-f. Four types of this kind of transistors were fabricated, 1) transistors with one sinusoidal edge and the other edge smooth (figure 6.8c), and transistors with two sinusoidally shaped edges that 2) are in phase (figure 6.8d), 3) have 90° phase difference (figure 6.8e), and 4) have 180° phase difference (figure 6.8f). As reference, also normal transistors were available (figure 6.8a). The transistors are made



(a) medium extra rough



(b) strong extra rough

Figure 6.9. Autocovariance functions of the LWR of transistors with medium extra rough gates and with strong extra rough gates. •) Based on the extracted gate edges. ■) Compensated for the smoothing out of the LWR due to diffusion of the extensions.

in the 130 nm process, that was described earlier. The examined average gate lengths range from 50 nm to 100 nm, while the gate width was fixed at 1.0 μm. Only NMOS transistors were available. In the first experiment, the amplitude of the sinusoidally shaped edges of the gates was varied from 0 % to 40 % of the average gate length, while the period was kept constant at 120 nm. In the second experiment, the period was varied from 40 nm to 1000 nm at a constant amplitude of 20 % of the gate length. A SEM picture was made of each of the fabricated transistors from which the exact line shapes were extracted. This information is required later to calculate the expected drain current of the transistors. Besides transistors with sinusoidally shaped gate edges, also transistors with extra rough edges have been fabricated, as displayed in figure 6.8b. This was achieved by randomly varying the e-beam dose along the edges of the transistor. Two varieties of extra rough transistors were measured, one with medium extra-rough gates and the other with strong extra-rough gates. The respective autocovariance functions of the LWR are displayed in figure 6.9. Per transistor length 65 device pairs were available.

6.3.2 Sinusoidally-shaped gate edges

In this subsection the experimental results with respect to the sinusoidally shaped transistors are compared with calculations. As a first approximation, the current is calculated by:

$$I_D = \frac{1}{W} \int_W I_{D,noLWR}(L_{local}(z))dz, \tag{6.20}$$

where $I_{D,noLWR}(L_{local}(z))$ is the drain current of an ideal transistor without LWR with a gate length of $L_{local}(z)$. This ideal current is measured on the reference transistors, that approximately have straight edges. Only a limited discrete set of reference gate lengths is available, and an interpolation algorithm is used to determine $I_{D,noLWR}$ for all available local gate lengths.

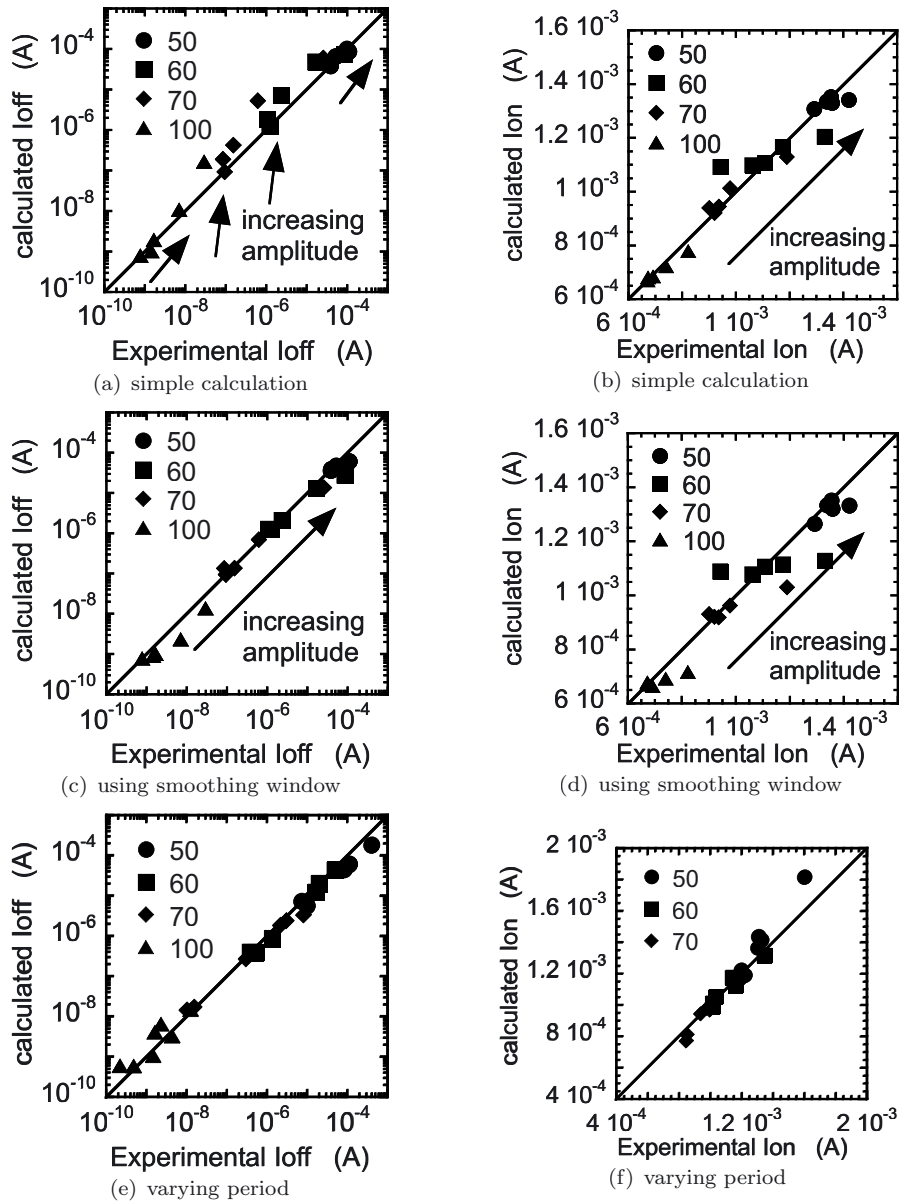
Figures 6.10a+b compare the calculated off- and on-state currents with the experimental data for the transistors of which one edge has the sinusoidal shape. These figures should be read as follows. On the x-axis the experimentally obtained current is plotted and on the y-axis the calculated current. Each symbol represents one transistor and the shape of the symbol is related to its average length. For each average length results are shown for amplitudes of the sinus of 0 % (reference), 5 %, 10 %, 20 % and 40 % of the gate length. The amplitude increases in the direction of the arrows.

For the off-state current (figure 6.10a) it is observed that (6.20) overestimates the experimentally observed current for increasing amplitudes. This can be explained by the fact that, by using (6.20), we have implicitly assumed that the tips of the extension regions exactly follow the gate, as is schematically displayed in figure 6.11a. However, in reality the roughness is smoothed out because of diffusion of the extensions during the processing after their implantation. This results in the situation displayed in figure 6.11b. Mathematically this can be taken into account by replacing L_{local} in (6.20) by L_{smooth}, which is given by:

$$L_{smooth}(z) = \frac{1}{2W_{smooth}} \int_{z-W_{smooth}}^{z+W_{smooth}} L_{local}(z')dz', \qquad (6.21)$$

where $2W_{smooth}$ is the width of the applied rectangular smoothing window. Figure 6.10c compares the experimental results with the new calculations in which W_{smooth} was used as a fitting parameter and found to be equal to 30 nm. A reasonable agreement between calculation and experiment is observed. Note that the value of 30 nm is comparable to the junction depth of the extensions. Therefore it can be considered a realistic value.

For the on-state current (figure 6.10b) it is seen that the calculation gives a reasonable prediction of the experimental results, except for some extreme points. Applying the smoothing window does not significantly change the picture, as can be seen in figure 6.10d. This demonstrates that in strong inversion the effect of varying gate length can indeed be linearized to first order. The smoothing reduces the increase in the current for the shorter parts of the channel, while it also reduces the decrease in current for the longer parts. In the linear approximation



(a) simple calculation (b) simple calculation

(c) using smoothing window (d) using smoothing window

(e) varying period (f) varying period

Figure 6.10. Comparison of calculated and experimental off-state currents (a+c+e) and on-state currents (b+d+f) for the transistors with one sinusoidal gate edge. The average length of the transistors is given in the legends and the arrows indicate increasing amplitude from 0 % to 40 % of the gate length (a-d). a+b) Equation 6.20 is used in the calculation. c+d) The smoothed length (6.21) is used in the calculation. e+f) The amplitude is fixed at 20 % of the gate length and the period is varied.

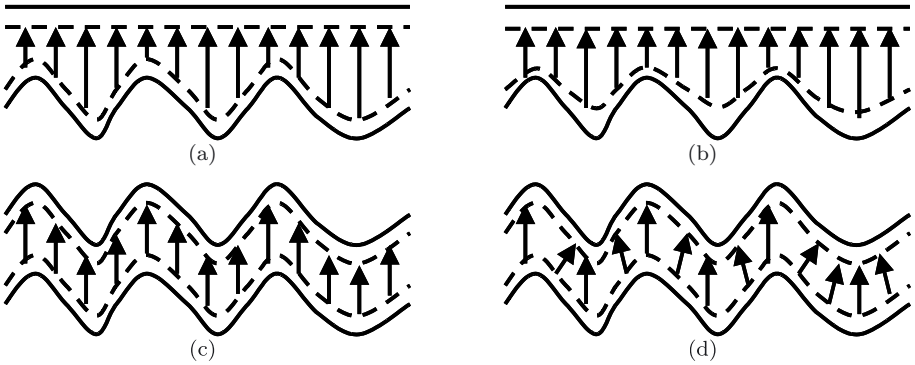
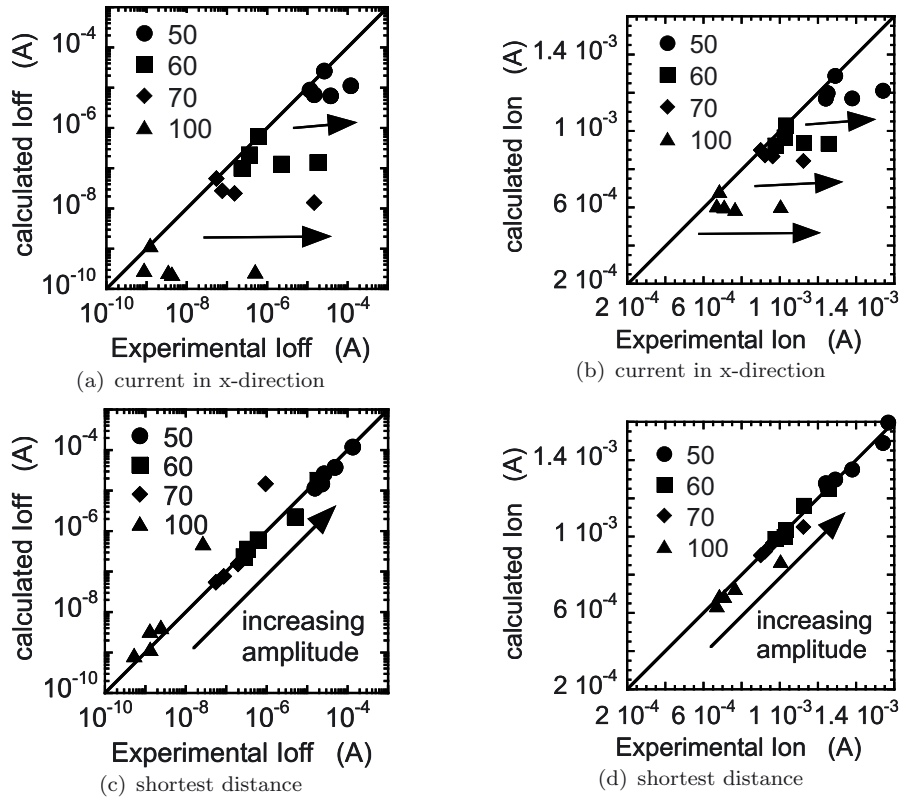


Figure 6.11. Schematic top-view drawings of transistors with one sinusoidal gate edge (a+b) and two sinusoidal gate edges in phase (c+d). The full lines represent the gate edge and the dashed lines the tips of the source and drain regions. The arrows denote the directions of the current flows.

these two effects exactly compensate each other.
Figures 6.10e+f compare the calculations of the off- and on-state current with experimental data for the case where the amplitude is fixed and the period is varied. The smoothing window is applied. A good agreement between experiment and theory is observed.
Now consider the case where both gate edges are sinusoidally shaped and in phase. Figures 6.12a+b compare the calculations of the off- and on-state current with experimental data for the case where the period is fixed and the amplitude is varied. The smoothed gate length was used in the calculations. It is observed that the current is seriously underestimated. This is caused by the assumption that the current only flows in the x-direction, as is schematically displayed in figure 6.11c. Since both edges are in phase, the gate length does not vary and the calculated current is independent of the amplitude. However, in reality the current mainly flows in the direction of the shortest distance between source and drain, as is shown in figure 6.11d. When the length in (6.20) is replaced with this shortest distance, while still taking into account the smoothing of (6.21), the calculation is found to accurately describe the experimental data, as is shown in figures 6.12c+d.
Finally, figure 6.13 shows the results for the cases where the two edges have a phase difference of 90° and of 180°. In these cases, to avoid too short transistors, the amplitude per edge is varied from 0 % to 20 %, instead of 40 %. The experimental data are seen to be well described. In the calculation a smoothing window with $W_{smooth} = 30$ nm is applied



(a) current in x-direction

(b) current in x-direction

(c) shortest distance

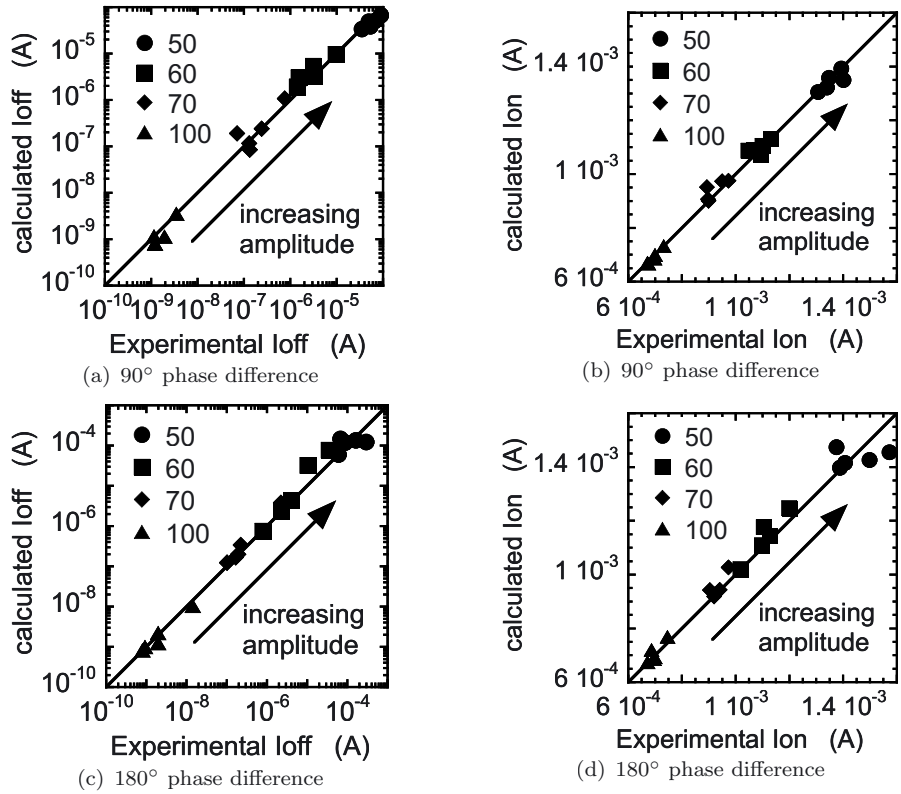
(d) shortest distance

Figure 6.12. Comparison of calculated and experimental off-state current (a+c) and on-state current (b+d) for the transistors with two sinusoidal gate edges in phase. The average length of the transistors is given in the legends and the arrows indicate increasing amplitude from 0 % to 40 % of the gate length. a+b) The current is assumed to flow in the x-direction c+d) The current is assumed to flow in the direction of the shortest distance between source and drain.

and the current is assumed to flow in the shortest direction from source to drain.

6.3.3 Extra rough gates

To evaluate the impact that the diffusion of the extension regions has on the LWR of the extra rough gates, (6.21) is applied to the extracted local gate lengths. From this, new autocovariance functions ($R_{LWR,smooth}$) are extracted, that are compared to the original autocovariance functions in figure 6.9. It is observed that $R(0) \equiv \sigma^2_{LWR}$ decreases. This means that the increase in off-state current due to LWR decreases because of smoothing. However, together with the decrease in σ^2_{LWR}, W_c is seen to



(a) 90° phase difference

(b) 90° phase difference

(c) 180° phase difference

(d) 180° phase difference

Figure 6.13. Comparison of calculated and experimental off-state current (a+c) and on-state current (b+d) for the transistors with two sinusoidal gate edges with a phase difference of 90° (a+b) and 180° (c+d). The average length of the transistors is given in the legends and the arrows indicate increasing amplitude from 0 % to 20 % of the gate length.

increase in such a way that the area under the autocovariance functions remains unchanged. In other words, smoothing out of the roughness does not change the magnitude of the parameter fluctuations.
Mathematically the impact of smoothing on the autocovariance function can be calculated as follows:

$$R_{LWR,smooth}(d) = [SW * SW * R_{LWR}](d), \tag{6.22}$$

where the smoothing window $SW(z) = 1/2W_{smooth}$ for $|z| < W_{smooth}$ and $SW(z) = 0$ for $|z| > W_{smooth}$. Since the area under $SW(z)$ is equal to one, it immediately follows that smoothing indeed does not change the area under the autocovariance function.
Applying (6.22) to the autocovariance function of the first-order autore-

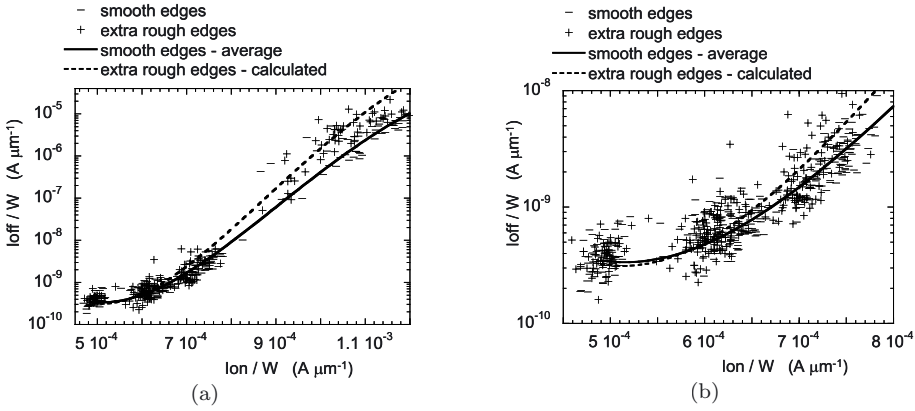


Figure 6.14. $I_{on} - I_{off}$ curves for transistors without extra roughness and for transistors with strong extra roughness. Symbols represent experimental data and the full line gives the average for the transistors with no extra roughness. The dashed line is the calculated curve for the extra rough transistors. The full $I_{on} - I_{off}$ curve is shown (a) and also the first part of the curve is shown (b).

gressive process (6.4) gives:

$$\sigma^2_{LWR,smooth} = \frac{W_c}{2W_{smooth}} \left(1 - \frac{W_c}{4W_{smooth}} \left(1 - e^{-\frac{4W_{smooth}}{W_c}}\right)\right) \cdot \sigma^2_{LWR}, \tag{6.23}$$

where $\sigma^2_{LWR,smooth} = R_{LWR,smooth}(0)$. From this it follows that smoothing effectively reduces σ_{LWR} when $W_c < 4W_{smooth}$ and that it is more effective for smaller W_c. Reducing W_c also reduces the parameter fluctuations, as follows from (6.8) and (6.14), but it could also decreases yield, as follows from (6.18) and (6.19).

It is more difficult to calculate the impact of the fact that the current does not flow purely in the x-direction. Numerical evaluation of the extra rough lines revealed no significant changes in the autocovariance functions. However, the average gate length is found to be reduced by approximately 0.5 nm. This is considered to be insignificant and will be neglected.

We will now look at some experimental results. Figure 6.14 shows the $I_{on} - I_{off}$ curves for the transistors with no extra LWR and for the transistors with strong extra LWR. The symbols represent measurement data and the full line shows the average for the transistors without extra LWR. From this average, the $I_{on} - I_{off}$ curve for the transistors with extra rough gates is calculated using (6.11) with $\sigma_{LWR,smooth}$ (dashed line). Reasonable agreement with the experimental data is observed, but a lot of scatter is present on the data. In a future experiment it would be

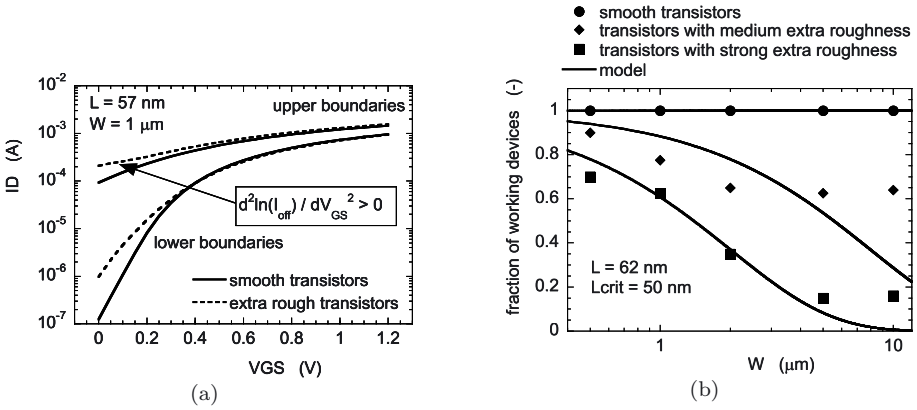


Figure 6.15. a) Upper and lower boundaries of the measured population of drain currents as a function of the gate bias. Results are shown for transistors with no extra roughness (full lines) and for transistors with strong extra roughness (dashed lines). b) Device yield as a function of the device width for transistors with no extra roughness (•), medium extra roughness (⧫), and strong extra roughness (■). Lines represent the model out of subsection 6.2.3.

advisable to increase the population size above the 65 device pairs per geometry that were available for this experiment. This would also allow the investigation of the expected increase in parameter fluctuations itself.

6.3.4 Yield

To illustrate the impact of LWR on yield, a criterion with respect to punch through is defined. Figure 6.15a shows the range of measured drain currents as a function of the gate bias for the population of transistors with no extra roughness and for the population of transistors with strong extra roughness. The gate length in this figure is equal to $L_{gate} = 57$ nm. In the presence of extra roughness, it is observed that the upper boundary of this range bends upwards at low values of the gate bias. This shows that the gate starts to loose its control over the channel and it is an indication of the onset of punch through. Therefore, as yield criterion we will say that a transistor fails when $d^2ln(I_{off})/dV_{GS}^2 > 0$ at $V_{GS} = 0$ V. Using this criterion, figure 6.15b shows the experimentally obtained device yield (symbols) as a function of the device width for transistors with an average gate length of $L_{gate} = 62$ nm. Results are shown for the case with no extra roughness, medium extra roughness and strong extra roughness. It is seen that the yield decreases with increasing LWR and with increasing device width. This is in agreement

with the analysis presented in subsection 6.2.3. The model out of this subsection is seen to be in reasonable agreement with the measurement data for $L_{critical} = 50$ nm.

6.4 Prediction of the impact of line-width roughness and scaling guidelines

After deriving and experimentally testing models regarding the impact of LWR, we will now use these models to predict the impact of LWR on future technologies. Based on this, requirements for LWR will be specified. In order to be able to make these predictions, it is necessary to know dV_T/dL and $dln(I_{off})/dL$. For this, 2D device simulations have been employed. Technology parameters are taken from the ITRS roadmap and are listed in table 6.1. The channel doping is assumed to be uniform, which means that the halos are fully overlapping. The doping level is chosen in such a way that the ITRS requirements regarding the on-state current are met. Both the high-performance and the low-power option are investigated. Also listed in table 6.1 are the simulation results regarding the threshold voltage, on-state current, off-state current, dV_T/dL and $dln(I_{off})/dL$. For the 130 nm technology node, it is observed that the simulated values of these parameters are approximately equal to their experimental counterparts, in case of the NMOS transistors (also see figure 6.4). This provides some confidence in the simulated results for future technologies.

Knowing the sensitivities of the threshold voltage and off-state current to the channel length, we can calculate the expected threshold voltage fluctuations and increase in off-state current, using the theory that was presented in section 6.2. In these calculations the LWR is assumed to be described by the first-order autoregressive process, displayed in figure 6.2. Smoothing out of the roughness because of diffusion of the extension regions has not been taken into account, because this effect is expected to decrease for future technologies.

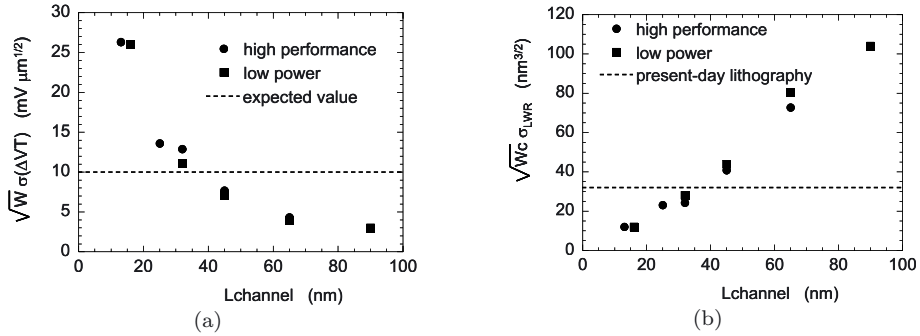
Figure 6.16a shows the expected threshold voltage mismatch (symbols) as a function of the channel lengths for the examined technology nodes. Also shown are the expected fluctuations, based on the scaling law that was presented in section 5.4 (dashed line). It is observed that, for a well optimized technology, LWR starts to become important for channel lengths below 40 nm. When the mismatch due to LWR is required to be smaller than the expected fluctuations without LWR, this gives a maximum to the allowed $\sqrt{W_c} \cdot \sigma_{LWR}$. This requirement is plotted in figure 6.16b. The dashed line represents the current status of what a gate-patterning process can achieve. When LWR does not decrease for future technologies and we wish to keep parameter fluctua-

Table 6.1. Input for and results of the 2D simulations, used to determine the sensitivity to the gate length of the threshold voltage at $V_{DS} = 50$ mV and the off-state current at $V_{DS} = V_{DD}$

node (nm)	130	90	65	45	32
			high performance		
V_{DD} (V)	1.2	1.0	0.9	0.7	0.5
$t_{ox,eff}$ (nm)	2.3	2.0	1.9	1.4	1.0
$L_{channel}$ (nm)	65	45	32	25	13
N_A (cm^{-3})	$1.3 \cdot 10^{18}$	$1.3 \cdot 10^{18}$	$1.5 \cdot 10^{18}$	$1.6 \cdot 10^{18}$	$2.3 \cdot 10^{18}$
V_T (V)	0.290	0.209	0.193	0.124	0.075
I_{on} ($\mu A \mu m^{-1}$)	869	915	913	924	897
I_{off} ($A \mu m^{-1}$)	$8.5 \cdot 10^{-9}$	$4.0 \cdot 10^{-7}$	$2.6 \cdot 10^{-6}$	$7.2 \cdot 10^{-6}$	$3.0 \cdot 10^{-5}$
$\frac{dV_T}{dL}$ ($mV nm^{-1}$)	3.05	5.42	9.09	9.56	18.5
$\frac{dln(I_{off})}{dL}$ (%nm^{-1})	-23.7	-32.6	-43.4	-42.3	-56.2
			low power		
V_{DD} (V)	1.2	1.1	1.0	0.9	0.7
$t_{ox,eff}$ (nm)	3.0	2.6	2.2	1.7	1.4
$L_{channel}$ (nm)	90	65	45	32	16
N_A (cm^{-3})	$0.8 \cdot 10^{18}$	$1.1 \cdot 10^{18}$	$1.7 \cdot 10^{18}$	$2.8 \cdot 10^{18}$	$3.4 \cdot 10^{18}$
V_T (V)	0.320	0.329	0.359	0.375	0.309
I_{on} ($\mu A \mu m^{-1}$)	600	619	607	612	599
I_{off} ($A \mu m^{-1}$)	$9.8 \cdot 10^{-10}$	$2.5 \cdot 10^{-9}$	$3.3 \cdot 10^{-9}$	$2.8 \cdot 10^{-9}$	$1.5 \cdot 10^{-7}$
$\frac{dV_T}{dL}$ ($mV nm^{-1}$)	2.13	2.74	5.04	7.84	18.3
$\frac{dln(I_{off})}{dL}$ (%nm^{-1})	-13.7	-21.5	-33.8	-45.8	-86.4

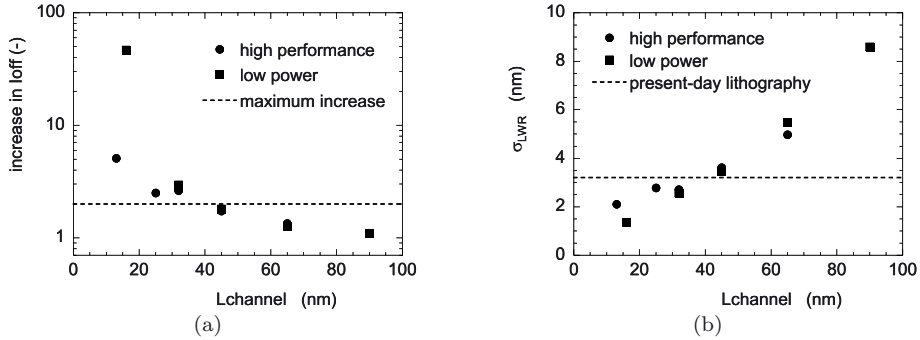
tions under control, the transistor needs to optimized in such a way that $dV_T/dL < 7$ $mV nm^{-1}$. More generally, it can be stated that the magnitude of the LWR (partly) determines the size of the design space for MOSFET development.

Now consider the off-state current. Figure 6.17a shows the expected increase in off-state current as a function of the channel length. It is observed that below channel lengths of 40 nm this increase is larger than a factor 2. If we require the increase to be smaller, this results in a maximum allowed σ_{LWR}. This requirement is plotted in figure 6.17b. When LWR does not decrease for future technologies and we wish to keep the increase in off-state current under control, the transistor needs to optimized in such a way that $dln(I_{off})/dL < 37$ %nm^{-1}. Again it



(a) (b)

Figure 6.16. a) Predicted mismatch in the threshold voltage due to LWR (symbols) as a function of the channel length. The dashed line gives the expected overall mismatch in the threshold voltage. b) Requirement on the LWR to keep the threshold-voltage mismatch below the overall expected value. The to the channel lengths corresponding technology nodes are listed in table 6.1.



(a) (b)

Figure 6.17. a) Predicted increase in the off-state current due to LWR (symbols) as a function of the channel length. b) Requirement on the LWR to keep the increase below a factor 2. The technology nodes corresponding to the channel lengths are listed in table 6.1.

can be stated that the magnitude of the LWR (partly) determines the size of the design space for MOSFET development.
Note that the requirement shown in figure 6.17b is somewhat more relaxed than the ITRS roadmap requirement, which is based on yield. In the ITRS roadmap no mention is made about the correlation width W_c. Therefore, in addition to the ITRS roadmap, we propose a new figure of merit for LWR, namely $\sqrt{W_c}\sigma_{LWR}$ for which the requirements are given in figure 6.16b.

6.5 Conclusions

This chapter investigated the impact of line-width roughness on MOSFET mismatch, off-state current and yield. The LWR was described by a first-order autoregressive process, which is represented by an autocovariance function. The magnitude of this autocovariance function was found to be equal to σ^2_{LWR} and its width equal to W_c.

It was reasoned that line-width roughness (LWR) affects MOSFET parameters through their dependence on the gate length. In the calculation of the impact of LWR on threshold-voltage mismatch all equations were linearized. The variance followed from standard matching theory and was calculated by twice taking the convolution of the autocovariance function with the geometry function and by multiplying the result with the squared sensitivity of the threshold voltage to the gate length. For wide transistors the variance was found to be inversely proportional to the transistor width, which is the one dimensional equivalent of the one-over-area model presented earlier. For very narrow transistors the gate length does not have the space to locally vary within the device and the variance in average length is equal to σ^2_{LWR}. By comparing the theoretically calculated fluctuations due to LWR to the mismatch of a 130 nm technology, it was found that LWR does not give a significant contribution to the parameter fluctuations for gate-lengths ranging down to 80 nm.

Locally the off-state current was reasoned to possess a log-normal distribution. The strong asymmetry of this distribution causes the LWR to increase the average off-state current. On a 130 nm technology this increase was still found to be small.

The impact of LWR on yield was calculated by first evaluating the probability that locally a device has a shorter length than a certain critical gate length. The yield followed from the amount of times that this probability appears in a circuit. This resulted in the requirement that $6\sigma_{LWR} < 0.3L_{gate}$, which is the same as the ITRS roadmap requirement on LWR.

Transistors with sinusoidal gate shapes were fabricated in order to experimentally evaluate the averaging processes of the local properties of the LWR. It was found that diffusion of the extension regions smoothes out the roughness. This was taken into account by applying a smoothing window to the LWR with a width of two times 30 nm. This smoothing results in a reduced increase in the off-state current due to LWR, but parameter fluctuations remain unchanged. It was also observed that the current mainly flows in the direction of the shortest distance between source and drain, but this has little impact on the more realistic situation where the roughness is random.

Besides transistors with sinusoidally shaped gates, also transistors with extra rough gates were created. The developed models with respect to the increase in off-state current and yield were validated, but the experimental accuracy was low.
Using the developed models, predictions were made regarding the threshold voltage mismatch and increase in off-state current caused by LWR. It is concluded that these effects start to play a role for technologies for which the nominal transistor has a channel length smaller than 40 nm. Requirements on the LWR were presented to keep parameter fluctuations and increase in off-state current under control. This resulted in a new figure of merit that also takes into account the correlation width of the LWR.

Chapter 7

CONCLUSIONS, FUTURE WORK AND OUTLOOK

7.1 Conclusions

In this work we have addressed the matching properties of deep submicron MOSFETs. In five chapters we have treated the modeling of the mismatch in the drain current, mismatch parameter extraction, the physical origins of mismatch, technological aspects and the impact of line-edge roughness. This includes all major areas of study related to MOSFET mismatch at the device level. The overall conclusions are presented chapter by chapter. The emphasis lies on the original contributions made by this work. For more extensive conclusions we refer to the corresponding chapters.

Chapter 2: Measurement and modeling of mismatch in the drain current. A physics-based deep-submicron model to describe the mismatch in the drain current has been developed and for the first time demonstrated on a 180 nm technology. As opposed to literature, we model the impact of a mismatch in the threshold voltage and a mismatch in the current factor separately. This results in a continuous model that is valid from moderate to strong inversion and in as well the linear as the saturation regime. The inaccuracy is smaller than 20 % at all bias conditions above threshold.

Chapter 3: Parameter extraction. The most common methods to extract the mismatch in threshold voltage and current factor are, for the first time, directly compared. Significant differences are observed, which can seriously affect the conclusions with respect to the matching performance of a technology. The differences between methods are related to small modeling errors or the nonexistence of a proper model for the weak inversion regime. The preferred method depends on the

application. With respect to model and measurement accuracy, current-mismatch fitting-methods yield the best results. A disadvantage of these methods is that they are slow. Applying a current criterion is much faster and also yields excellent measurement accuracy. However, the physical content of the extracted threshold voltage mismatch is less well defined, and it is difficult to use it to characterize a technology. The maximum slope method is reasonably fast, provides understandable results, but it is sensitive to the contact resistance, which can lead to inaccuracies. The three- and four-points methods are most sensitive to noise introduced by the measurement setup. However, they are very fast and provide understandable results. When the three- or four-points method is applied it is required to use fixed gate-overdrive voltages for the bias points.

Chapter 4: Physical origins of MOSFET mismatch. By solving the current equations we find that the most commonly applied $1/\sqrt{area}$ law for mismatch is only valid in strong inversion at low values of the drain bias. In addition to literature, we have found that in weak inversion deviations are mainly caused by an exponentially larger contribution to the mismatch of sidewall transistors. In strong inversion at higher values of the drain bias, we find that the lateral non-uniformity of the inversion layer causes a logarithmic, i.e. weak, deviation, which is in accordance with recent literature. We reason that this non-uniformity also results in asymmetry of the MOSFET.
The impact on the mismatch of doping fluctuations in the channel and gate, mismatch in the oxide charge and mismatch in surface roughness scattering is calculated. Besides the direct impact of doping fluctuations on the threshold voltage, we derived that Coulomb scattering plays a significant role. As opposed to literature, we reason that fluctuations in Coulomb scattering appear as apparent fluctuations in the threshold voltage. By this, we can explain a large part of the gap between the calculated and experimentally obtained mismatch in the threshold voltage. Experimental testing of our model shows excellent descriptive and reasonable predictive behavior.

Chapter 5: Technological aspects. Examples of the impact of technological parameters on the matching behavior have been presented. It is confirmed that the granular structure of the poly-silicon gate has a significant impact. As new technological issue, we find that halos can unintentionally be implanted through the gate. This results in a serious degradation of the matching performance by causing extra fluctuations in either the channel doping or gate depletion.
The scaling of the matching performance of technologies has been addressed. It was concluded that the matching performance improves beyond what is expected from basic scaling laws. However, it was also

reasoned that research efforts remain necessary to keep parameter fluctuations under control. This is especially important for the minimum sized transistor, for which the parameter fluctuations get worse as dimensions are scaled down.

Chapter 6: Impact of line-edge roughness on parameter fluctuations, off-state current and yield. When we started working on line-edge roughness (LER), this was a relatively new subject. Therefore, most of the presented work is original. In order to evaluate the impact of LER, we developed a method to characterize the roughness itself, we derived models to calculate the impact of LER on transistor behavior, we experimentally tested these models and we predicted the impact of LER for future technologies.
Edge roughness of the gate is described by a first order autoregressive process. It is characterized by the standard deviation of the roughness and a correlation width. The impact of LER on parameter fluctuations, increase in off-state current and yield is calculated. Compared to experimental values for parameter fluctuations, line-edge roughness is not expected to have a significant impact down to the minimum measured gate length of 80 nm. Developed models are verified on specially fabricated transistors with sinusoidally-shaped gate-edges. Based on these models and device simulations, it is predicted that line-edge roughness will start to become important for devices with 32 nm channel lengths for modern-day gate-patterning processes. Based on the standard deviation and the correlation width of the LER, a new figure of merit has been introduced to describe the impact of LER on parameter fluctuations.

7.2 Future work

The variability of the minimum sized transistor increases with the down-scaling of transistor dimensions. Good matching performance has always been a technology requirement for analog applications, and it has also become a necessity for digital designs. We end this book by introducing seven possible topics of future research regarding the stochastic properties of MOSFETs and by presenting an outlook.

Develop one model for the mismatch in the drain current for the complete inversion regime. It was observed that the mismatch in the weak inversion regime cannot be predicted from the mismatch in strong inversion, because of the effect of the isolation and because of halos. As supply voltages scale down, analog design is pushed more and more into the weak inversion region. Therefore, it is necessary to develop a model that is valid in the whole inversion regime, i.e. weak, moderate and strong inversion. This could be a model that divides the transistor in six sub-transistors, as displayed in figure 7.1. Half of these sub-transistors

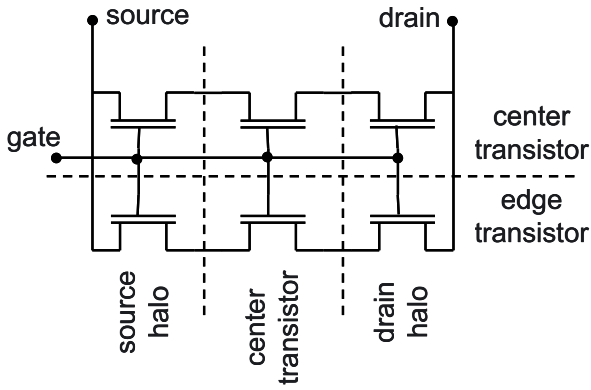


Figure 7.1. Proposed model to describe the mismatch in the drain current in the complete inversion regime

model the edge transistors, while the other half model the center transistor; four of the sub-transistors model the halos at the source and drain sides, while two sub-transistors model the center. The mismatch of each of the sub-transistors would be modeled separately by the mismatch model provided in chapter 2.

Provide more physical models for the mismatch in the mobility. It was reasoned in chapter 4 that mismatch due to Coulomb scattering can give a significant contribution to the overall mismatch. However, this analysis was based on a semi-empirical model for the mobility. It would be of great interest to use more physical models to derive the stochastic properties of the mobility. These models should take the full geometry of the electrostatic problem into account.

Evaluate the mismatch in the gate tunnelling current. As the thickness of the gate oxide scales down, the gate tunnelling current becomes significant. This changes the MOSFET characteristics and requires adaptation of models and extraction routines. The tunnelling current itself is strongly dependent on the oxide thickness. Therefore, it is by itself susceptible to stochastic variations that need to be studied.

Evaluate the mismatch in devices with high-k dielectrics. To get rid of large tunnelling currents, high-k materials are foreseen as gate dielectric. Besides changing the effective oxide thickness, using a different dielectric will give rise to different values of dielectric charge, a different concentration of interface states and it can significantly influence mobility. All these effects will have an impact on the matching behavior of a technology, which needs to be studied before a new dielectric is introduced.

Matching properties of new devices. In order to reach sufficient performance at ultra-small dimensions, alternative device concepts are introduced, as discussed at the end of chapter 5. These new concepts will give rise to new sources of fluctuations, which need to be examined. As an example consider the FinFET. The lower boundary to the stochastic fluctuations of such a device are foreseen to be caused by fin-width and fin-length roughness. Their impact can be examined by following the same approach and defining similar experiments as was done in chapter 6 to evaluate the impact of line-edge roughness.

Investigate the matching properties of MOSFETs at the circuit level. In this book the matching properties of MOSFETs were investigated by looking at matched transistor pairs. In reality a MOSFET is operated in a circuit environment. This gives rise to asymmetries that can cause, often unforeseen, systematic contributions to the mismatch. These need to be investigated by a proper set of test structures in order to define a set of layout rules. Furthermore, other circuit elements, such as interconnects, can add to the variability. As an example, consider the following experiment. As test structures one could design a matched transistor pair (NMOS and PMOS), a matched invertor, and an SRAM cell. In each of these structures the dimensions of the transistors are kept the same. From the matching properties of the matched transistor pair one should be able to predict the matching properties of the invertor from which one should be able to predict the symmetry of the SRAM cell. Other circuit elements can cause deviations from this expected behavior, which can now be evaluated.

Investigate the small-signal matching properties of MOSFETs. In matching analysis only the DC behavior of MOSFETs is considered. Small signal parameters are generally ignored, while it would be very interesting to measure the matching properties of e.g. the overlap capacitance. However, measurement accuracy is always an issue in matching analysis. Therefore, this kind of research probably requires dedicated test structures.

7.3 Outlook

Variability is but one of many scaling issues. Nevertheless, the stochastic properties of MOSFETs will become a limiting factor. This means that a significant effort remains necessary to keep variabilities under control and to try to decrease them. Using retrograde doping profiles improves the matching performance, but it seriously complicates the process. Fully depleted SOI devices and FinFETs have the potential for very good matching performance, but this has yet to be demonstrated, while these architectures also contain new possible sources of parameter

fluctuations.
In general, variability increases when device dimensions scale down, and the point has been reached where a digital transistor cannot anymore be considered as fully digital with 100 % certainty. For instance, MOSFET mismatch could result in significant variability in the MOSFET delay, which could cause timing issues. This requires digital design methodologies to take variabilities into account, as has always been the case for analog design. Based on new design methodologies, new figures of merit for device optimization could be derived. For instance, at some point it might be more efficient for a certain application to have somewhat more accurate transistors instead of very fast ones. More generally, it is predicted that device optimization will become part of the circuit-design methodology. This kind of optimization would be strongly application dependent, but it can potentially lead to better performing circuits.

Appendix A
List of symbols

Symbol	Unit	Description
$A_{0,\Delta P}$		proportionality constant related to the $\sigma_{\Delta P} \propto 1/\sqrt{WL}$ law
$A_{0,\Delta P_1,\Delta P_2}$, $A_{L,\Delta P_1,\Delta P_2}$,.. ..,$A_{W,\Delta P_1,\Delta P_2}$, $A_{WL,\Delta P_1,\Delta P_2}$		parameters decribing the width and length dependence of the correlation between ΔP_1 and ΔP_2
$A_{L,\Delta P}$		parameter used to describe the deviation from the $\sigma_{\Delta P} \propto 1/\sqrt{WL}$ law for short devices
a_{ph}	$\mathrm{cm^2 s^{-1}}$	parameter describing the gate-bias dependence of the mobility due to phonon scattering
a_{sr}	$\mathrm{cm^2 s^{-1}}$	parameter describing the gate-bias dependence of the mobility due to surface-roughness scattering
$A_{W,\Delta P}$		parameter used to describe the deviation from the $\sigma_{\Delta P} \propto 1/\sqrt{WL}$ law for narrow devices
$A_{WL,\Delta P}$		parameter used to describe the deviation from the $\sigma_{\Delta P} \propto 1/\sqrt{WL}$ law for short and narrow devices
B_{QM1}	$\mathrm{V^{1/3} cm^{2/3}}$	parameter used in the calculation of quantummechanical effects
B_{QM3}	$\mathrm{cm^{1/3} C^{-1/3}}$	parameter used in the calculation of quantummechanical effects
C_D	$\mathrm{F\ cm^{-2}}$	depletion-layer capacitance
C_{GC}	$\mathrm{F\ cm^{-2}}$	gate-to-channel capacitance
C_{ox}	$\mathrm{F\ cm^{-2}}$	oxide capacitance
d	cm	distance along line
D_{N_A}	cm	location of doping concentration peak
d_{QM}	cm	distance used to describe the increase in $\sigma_{\Delta V_{GS}}$ due to quantummechanical effects
E_{eff}	$\mathrm{V\ cm^{-1}}$	effective field
E_F	eV	Fermi energy
E_i	eV	intrinsic energy
$E_{N_A(y)}$	$\mathrm{V\ cm^{-1}}$	electrical field caused by the charge sheet $qN_A(y)dy$

E_{QM2}	V cm^{-1}	parameter used in the calculation of quantummechanical effects
E_s	V cm^{-1}	electrical field at the oxide-silicon interface
E_1	V cm^{-1}	electrical field at pinch-off point
$f_{\delta P}$	cm^2	normalized power spectrum of δP
G	cm^{-2} / cm^{-1}	geometry function
g_m	A V^{-1}	transconductance
g_{mmax}	A V^{-1}	maximum transconductance
g_{out}	A V^{-1}	output conductance
I_0	A	from weak inversion extrapolated drain current at $V_{GS} = V_T$
I_{cs}	A	current delivered by the current-source transistor
I_D	A	drain current
$I_{D,noLWR}$	A	drain current of a device without line-width roughness
I_{fs}	A	full-scale current
I_{off}	A	off-state current (or leakage current)
$I_{off,local}$	A cm^{-1}	local off-state current at a certain position z along the gate
J	A cm^{-2}	current density
k	J K^{-1}	Boltzmann's constant ($= 1.38 \cdot 10^{-23}$)
K_{sr}	V s^{-1}	parameter used in the calculation of μ_{sr}
K_{V_T}, K_δ		parameters used in the calculation of the impact of δV_T on the drain current
L	cm	channel length
$L_{channel}$	cm	metallurgical channel length
L_{crit}	cm	critical gate length
L_{eff}	cm	effective channel length
L_{gate}	cm	gate length
L_{local}	cm	local gate length
L_{mask}	cm	gate length on mask
L_{met}	cm	metallurgical channel length
l_p	cm	position of pinch-off point with respect to drain
L_{smooth}	cm	local gate length after applying a smoothing window
L_Δ	cm	correlation length of surface roughness
$l_{\delta P}$	cm	correlation length of the stochastic process describing δP
m_C		parameter used in the calculation of μ_C
n	cm^{-3}	electron concentration
n		$\equiv 1 + C_D/C_{ox}$
n		$\equiv \partial \ln(\mu)/\partial \ln(E_{eff})$
n_0	cm^{-3}	electron concentration if no microscopic fluctuations were present
N_A	cm^{-3}	doping concentration
N_{A0}	cm^{-3}	peak doping concentration
N_{BIT}		number of bits
N_{dev}		number of device pairs
N_{device}		number of devices
N_{dope}	cm^{-3}	doping concentration
N_f	cm^{-2}	fixed-oxide-charge density

n_i	cm^{-3}	intrinsic carrier concentration
N_p	cm^{-3}	doping concentration in gate
p	cm^{-3}	hole concentration
p	cm^{-1}	parameter in the second-order autoregressive model
P		a parameter
p_{device}		probability that a device fails
p_{local}		probability that locally the gate length is shorter than L_{crit}
p_μ		Fuchs scattering factor
q	C	elementary charge ($= 1.6 \cdot 10^{-19}$)
Q_{cs}	$\mathrm{C\ cm}^{-2}$	charge density in charge sheet
Q_D	$\mathrm{C\ cm}^{-2}$	depletion charge
Q_i	$\mathrm{C\ cm}^{-2}$	inversion-layer charge
q_p	C	point charge
Q_s	$\mathrm{C\ cm}^{-2}$	substrate charge
r	cm	radial direction
R	cm^2	autocovariance function related to the surface roughness
$R_{contact}$	Ω	contact resistance
R_D	$\Omega\ \mathrm{cm}^{-1}$	series resistance at the drain
$r'_{I_{off}}$		ratio of the off-state current in a realistic device incorporating microscopic fluctuations and an ideal device without these fluctuations
R_{LWR}	cm^2	autocovariance function related to the line-width roughness
$R_{LWR,smooth}$	cm^2	autocovariance function related to the smoothed out line-width roughness
$R_{r'I_{off}}$		autocovariance function related to local variation of $r'I_{off}$
R_s	$\Omega\ \mathrm{cm}^{-1}$	series resistance
R_S	$\Omega\ \mathrm{cm}^{-1}$	series resistance at the source
SW	cm^{-1}	smoothing window
T	K	temperature
t_{GD}	cm	thickness of depletion layer in gate
t_{ox}	cm	oxide thickness
t_{oxeff}	cm	effective oxide thickness
t_{QM}	cm	quantummechanical increase in oxide thickness
t_{QMGD}	cm	parameter used to describe the impact of fluctuations in t_{QM} on gate depletion
t_{QMQD}	cm	parameter used to describe the impact of fluctuations in t_{QM} on the depletion-layer charge
$t_{\mu GD}$	cm	parameter used to describe the impact of fluctuations in the effective field on gate depletion
$t_{\mu QD}$	cm	parameter used to describe the impact of fluctuations in the effective field on the depletion-layer charge
V_{BS}	V	bulk-to-source voltage
V_{CB}	V	channel-to-bulk voltage
V_{CS}	V	channel-to-source voltage

V_{DS}	V	drain-to-source voltage
V_{DSsat}	V	saturation voltage
V_{FB}	V	flat-band voltage
V_{GC}	V	gate-to-channel voltage
V_{GS}	V	gate-to-source voltage
V_{q_p}	V	impact of a point charge on surface potential
v_{sat}	cm s^{-1}	saturation velocity
V_T	V	threshold voltage
V_{T0}	V	threshold voltage at $V_{BS} = 0$ V
V_{T0}	V	threshold voltage at $V_{DS} = 0$ V
$V_{T,local}$	V cm^{-1}	local threshold voltage at a certain position z along the gate
V_{Tlw}	V	threshold voltage of a long and wide transistor
$V_{Tnarrow}$	V	threshold voltage of a narrow transistor
W	cm	channel width
W_c	cm	correlation width related to the line-width roughness
$W_{c,I_{off}}$	cm	correlation width related to the microscopic fluctuations in $I_{off,local}$
W_D	cm	depletion-layer width
W_D^{QM}	cm	depletion-layer width, calculated including quantummechanical effects
W_{middle}	cm	width of center transistor
W_{N_A}	cm	width of doping concentration peak
W_{narrow}	cm	width of parasitic edge transistor
W_{smooth}	cm	half of the width of the smoothing window
$w_{\delta P}$		sensitivity of the drain current to δP
x	cm	direction from source to drain
y	cm	direction perpendicular to oxide-silicon interface
y_{sc}	cm	depth of a charge sheet
z	cm	width direction
z_μ	cm	inversion layer thickness
α		fitting parameter, used to describe the V_{BS} dependence of $\sigma_{\Delta V_T}$
α_1	cm^{-1}	parameter in the first-order autoregressive model
α_2	cm^{-1}	parameter in the second-order autoregressive model
β	A V^{-2}	current factor
β_0	A V^{-2}	current factor, without gate or drain bias dependent effects taken into account
γ	V$^{1/2}$	body-effect coefficient
γ_{BH}		Brooks-Herring screening parameter
δ		$= C_D/C_{ox}$
Δ	cm	rms value of the surface rourghness
δf_n		$\equiv e^{q\delta\psi_s/KT}$
δP		microscopic deviation of a parameter from its typical value
ΔP		mismatch in a parameter ($= P_2 - P_1$)
$\overline{\Delta P}$		average of the mismatch in a parameter
$\Delta' P$		deviation of a parameter from the ideal case without microscopic fluctuations

ΔU_T	V	$\equiv -(V_{GS} - V_T)\Delta I_D / I_D$
ΔW_D^{QM}	cm	quantummechanical increase of the depletion layer width
Δx_s	cm	position, with respect to the drain, of maximum sensitivity of the drain current to δV_T
$\Delta \psi_s$	V	shift in surface-potential due to the short-channel effect
$\Delta \psi_s^{QM}$	V	shift in surface-potential due to quantummechanical effects
ϵ_{ox}	F cm^{-1}	permittivity of silicon dioxide ($= 3.45 \cdot 10^{-13}$)
ϵ_{si}	F cm^{-1}	permittivity of silicon ($= 1.04 \cdot 10^{-12}$)
ζ_{sat}	A V^{-1}	parameter describing the drain-bias dependence of the current factor
ζ_{sr}	A V^{-1}	parameter describing the gate-bias dependence of the current factor
η		parameter used in the calculation of the effective field
θ	V^{-1}	mobility reduction factor
θ_{sat}	V^{-1}	parameter describing the drain-bias dependence of the current factor
θ_{sr}	V^{-1}	parameter describing the gate-bias dependence of the current factor ($= \theta$)
θ_1	V^{-1}	first-order mobility reduction factor
θ_2	V^{-2}	second-order mobility reduction factor
κ		scaling coefficient
λ	cm	parameter related to the range of the short-channel effect
μ	cm^{2}V^{-1}s^{-1}	mobility
μ_B	cm^{2}V^{-1}s^{-1}	bulk mobility
μ_C	cm^{2}V^{-1}s^{-1}	mobility limited by Coulomb scattering
μ_{fc}	cm^{2}V^{-1}s^{-1}	mobility limited by fixed-oxide-charge scattering
μ_{sat}	cm^{2}V^{-1}s^{-1}	drain bias limited part of mobility
μ_{sr}	cm^{2}V^{-1}s^{-1}	mobility limited by surface-roughness scattering
μ_{sr}	cm^{2}V^{-1}s^{-1}	gate bias limited part of mobility
$\mu_{\Delta P}$		average of the mismatch in a parameter
ξ		parameter related to the effect of non-abrupt junctions on the short-channel effect
ρ		correlation factor
ρ	Ω cm^{-3}	resistivity
ρ_{LWR}		autocorrelation function of the line-width roughness
ρ_{repeat}		measurement repeatability
$\rho(\Delta P_1, \Delta P_2)$		correlation between the mismatch in P_1 and the mismatch in P_2
σ		standard deviation
σ_{LER}	cm	standard deviation of the line-edge position
σ_{ln}		$\equiv \lvert dln(I_{off})/dL \rvert \sigma_{LWR}$
σ_{LWR}	cm	standard deviation of the local line-width
$\sigma_{LWR,smooth}$	cm	standard deviation of the smoothed out local line-width
$\sigma_{\Delta P}$		standard deviation of the mismatch in a parameter

σ_σ		standard deviation of extracted standard deviation
ϕ_B	V	surface potential in strong inversion
ϕ_F	V	Fermi potential
ϕ_{MS}	V	work function of gate
ϕ_t	V	thermal voltage ($= kT/q$)
ψ	V	potential
ψ_s	V	surface potential
ψ_s^0	V	long-channel surface-potential
ω_r	cm^{-1}	spacial frequency in the 1/r direction
ω_x	cm^{-1}	spacial frequency in the 1/x direction
ω_z	cm^{-1}	spacial frequency in the 1/z direction
$[f_1 * f_2](x)$		$\equiv \int_{-\infty}^{\infty} f_1(x') \cdot f_2(x - x')dx'$ (convolution integral)

Appendix B
List of acronyms

Symbol	Description
2D	Two Dimensional
3D	Three Dimensional
CMOS	Complementary Metal-Oxide Semiconductor
DAC	(current-steering) Digital-to-Analog Converter
DIBL	Drain Induced Barrier Lowering
FD	Fully Depleted
HDD	Highly Doped Drain
INL	Integrated Non-Linearity error
ITRS	International Technology Roadmap for Semiconductors
LDD	Lowly Doped Drain
LER	Line-Edge Roughness
LWR	Line-Width Roughness
MOS	Metal-Oxide Semiconductor
MOSFET	Metal-Oxide Semiconductor Field-Effect Transistor
SCE	Short-Channel Effect
SEM	Scanning Electron Microscope
SIMS	Secondary Ion Mass Spectrometry
SMU	Source-Monitor Unit
SNM	Static Noise Margin
SOI	Silicon-On-Insulator
SRAM	Static Random Access Memory
STI	Shallow Trench Isolation

Appendix C
Publications by the author

Journal paper:

1 J.A. Croon, M. Rosmeulen, S. Decoutere, W. Sansen and H.E. Maes, "An Easy-to-Use Mismatch Model for the MOS Transistor," *IEEE Journal of Solid-State Circuits*, vol. 37, no. 8, pp. 1056–1064, 2002

Conference papers:

1 J.A. Croon, M. Rosmeulen, S. Van Huylenbroeck and S. Decoutere, "A General Model for MOS Transistor Matching," in *Proc. of the* 29^{th} *European Solid-State Device Research Conference*, pp. 464-467, 1999

2 J.A. Croon, M. Rosmeulen, S. Decoutere, W. Sansen and H.E. Maes, "A Simple and Accurate Deep Submicron Mismatch Model," in *Proc. of the* 30^{th} *European Solid-State Device Research Conference*, pp. 356-359, 2000

3 J.A. Croon, M. Rosmeulen, S. Decoutere, W. Sansen and H.E. Maes, "A simple characterization method for MOS transistor matching in deep submicron technologies," in *Proc. of the 2001 International Conference on Microelectronic Test Structures*, pp. 213-218, 2001

4 J.A. Croon, H.P. Tuinhout, R. Difrenza, J. Knol, A.J. Moonen, S. Decoutere, H.E. Maes and W. Sansen, "A comparison of extraction techniques for threshold voltage mismatch," in *Proc. of the 2002 International Conference on Microelectronic Test Structures*, pp. 235-240, 2002

5 J.A. Croon, E. Augendre, S. Decoutere, W. Sansen and H.E. Maes, "Influence of Doping Profile and Halo Implantation on the Threshold Voltage Mismatch of a 0.13 μm CMOS Technology," in *Proc. of the* 32^{nd} *European Solid-State Device Research Conference*, pp. 579-582, 2002

6 J.A. Croon, G. Storms, S. Winkelmeier, I. Pollentier, M. Ercken, S. Decoutere, W. Sansen and H.E. Maes, "Line Edge Roughness: Characterization, Modeling and Impact on Device Behavior," *International Electron Device Meeting 2002*, pp. 307-310, 2002

7 J.A. Croon, L.H.A. Leunissen, M. Jurczak, M. Benndorf, R. Rooyackers, K. Ronse, S. Decoutere, W. Sansen and H.E. Maes, "Experimental investigation of the impact of line-edge roughness on MOSFET performance and yield," in *Proc. of the* 33^{rd} *European Solid-State Device Research Conference*, pp. 227-230, 2003

8 J.A. Croon, S. Decoutere, W. Sansen, H.E. Maes "Physical modeling and prediction of the matching properties of MOSFETs," in *Proc. of the* 34^{th} *European Solid-State Device Research Conference*, pp. 193-196, 2004

Unrelated publications:

1 J. Croon, S. Biesemans, S. Kubicek, E. Simoen, K. De Meyer and C. Claeys, "Freeze-out effects on the characteristics of deep submicron Si nMOSFETs in the 77 K to 300 K range," in *Proc. of the* 4^{th} *Symposium on Low Temperature Electronics and High Temperature Superconductivity*, pp. 187-198, 1997

2 J.A. Croon, H.M. Borsboom and A.F. Mehlkopf, "Optimization Of Low Frequency Litz-Wire RF Coils," in *Proc. of the* 7^{th} *Scientific Meeting & Exhibition of the International Society for Magnetic Resonance in Medicine*, pp. 740, 1999

3 J.A. Croon, B. Kaczer, G.S. Lujan, S. Kubicek, G. Groeseneken, M. Meuris, "Experimental analysis of a Ge-HfO_2-TaN gate stack with a large amount of interface states," *Acc. for publ. in the proc. of the 2005 International Conference on Microelectronic Test Structures*, 2005

References

[1] C. McDonald, "Copy EXACTLY! A paradigm shift in technology transfer method," *Proc. of the 1997 IEEE Advanced Semiconductor Manufacturing Conference*, pp. 414–417, 1997.

[2] H. Tuinhout, A. Bretveld, and W. Peters, "Current mirror test structures for studying adjacent layout effects on systematic transistor mismatch," *Proc. of the 2003 International Conference on Microelectronic Test Structures*, pp. 221–226, 2003.

[3] R. Keyes, "The effect of randomness in the distribution of impurity atoms on FET thresholds," *Applied Physics*, vol. 8, pp. 251–259, 1975.

[4] P. Kinget and M. Steyaert, "Impact of transistor mismatch on the speed-accuracy-power trade-off of analog CMOS circuits," *Proc. of the 1996 Custom Integrated Circuits Conference*, pp. 333–336, 1996.

[5] M. Pelgrom, A. Duinmaijer, and A. Welbers, "Matching properties of MOS transistors," *IEEE Journal of Solid-State Circuits*, vol. 24, no. 4, pp. 1433–1440, 1989.

[6] M. Pelgrom, H. Tuinhout, and M. Vertregt, "Transistor matching in analog CMOS applications," *International Electron Devices Meeting 1998*, pp. 915–918, 1998.

[7] A. Bhavnagarwala, X. Tang, and J. Meindl, "The impact of intrinsic device fluctuations on CMOS SRAM cell stability," *IEEE Journal of Solid-State Circuits*, vol. 36, no. 4, pp. 658–665, 2001.

[8] P. Stolk, H. Tuinhout, R. Duffy, E. Augendre, L. Bellefroid, M. Bolt, J. Croon, C. Dachs, F. Huisman, A. Moonen, Y. Ponomarev, R. Roes, M. Da Rold, E. Seevinck, K. Sreerambhatla, R. Surdeanu, R. Velghe, M. Vertregt, M. Webster, N. van Winkelhoff, and A. Zegers-Van Duijnhoven, "CMOS device optimization for mixed-signal technologies," *International Electron Devices Meeting 2001*, pp. 215–218, 2001.

[9] R. Van Overstraeten, G. Declerck, and G. Broux, "The influence of surface potential fluctuations on the operation of the MOS transistor in weak inversion,"

IEEE Transactions on Electron Devices, vol. ED-20, no. 12, pp. 1154–1158, 1973.

[10] K. Takeuchi, T. Tatsumi, and A. Furukawa, "Channel engineering for the reduction of random-dopant-placement-induced threshold voltage fluctuations," *International Electron Devices Meeting 1997*, pp. 841–844, 1997.

[11] P. Stolk, F. Widdershoven, and D. Klaassen, "Modeling statistical dopant fluctuations in MOS transistors," *IEEE Transactions on Electron Devices*, vol. 45, no. 9, pp. 1960–1971, 1998.

[12] A. Asenov, "Random dopant induced threshold voltage lowering and fluctuations in sub-0.1 μm MOSFET's: A 3-d "atomistic" simulation study," *IEEE Transactions on Electron Devices*, vol. 45, no. 12, pp. 2505–2513, 1998.

[13] T. Mizuno and A. Toriumi, "Experimental evidence for statistical-inhomogeneous distributed dopant atoms in Si metal-oxide-semiconductor field-effect transistor," *Journal of Applied Physics*, vol. 77, no. 7, pp. 3538–3540, 1995.

[14] H. Tuinhout, F. Widdershoven, P. Stolk, J. Schmitz, B. Dirks, K. van der Tak, P. Bancken, and J. Politiek, "Impact of ion implantation statistics on V_T fluctuations in MOSFETs: Comparison between decaborane and boron channel implants," *Proc. of 2000 Symposium on VLSI Technology*, pp. 134–135, 2000.

[15] E. Vittoz, "The design of high-performance analog circuits on digital CMOS chips," *IEEE Journal of Solid-State Circuits*, vol. SC-20, no. 3, pp. 657–665, 1985.

[16] K. Lakshmikumar, R. Hadaway, and M. Copeland, "Characterization and modeling of mismatch in MOS transistors for precision analog design," *IEEE Journal of Solid-State Circuits*, vol. SC-21, no. 6, pp. 1057–1066, 1986.

[17] J. Bastos, M. Steyaert, B. Graindourze, and W. Sansen, "Matching of MOS transistors with different layout styles," *Proc. of the 1996 International Conference on Microelectronic Test Structures*, pp. 17–18, 1996.

[18] T. Serrano-Gotarredona and B. Linares-Barranco, "A new five-parameter MOS transistor mismatch model," *IEEE Electron Device Letters*, vol. 21, no. 1, pp. 37–39, 2000.

[19] P. Drennan and C. McAndrew, "A comprehensive MOSFET mismatch model," *International Electron Devices Meeting 1999*, pp. 167–170, 1999.

[20] H. Tuinhout, M. Pelgrom, R. de Vries, and M. Vertregt, "Effects of metal coverage on MOSFET matching," *International Electron Devices Meeting 1996*, pp. 735–738, 1996.

[21] H. Tuinhout, A. Montree, J. Schmitz, and P. Stolk, "Effects of gate depletion and boron penetration on matching of deep submicron CMOS transistors," *International Electron Devices Meeting 1997*, pp. 631–634, 1997.

[22] R. Difrenza, P. Llinares, G. Ghibaudo, E. Robillart, and E. Granger, "Dependence of channel width and length on MOSFET matching for 0.18 μm

CMOS technology," *Proc. of the 30th European Solid-State Device Research Conference*, pp. 584–587, 2000.

[23] J. Bastos, *Characterization of MOS transistor mismatch for analog design*. K.U. Leuven, 1998.

[24] R. Difrenza, *Impact des fluctuations technologiques sur l'appariement du transistor MOS des filières 0.18 et 0.12µm*. Institut National Polytechnique de Grenoble, 2002.

[25] H. Yang, V. Macary, J. Huber, W.-G. Min, B. Baird, and J. Zuo, "Current mismatch due to local dopant fluctuations in MOSFET channel," *IEEE Transactions on Electron Devices*, vol. 50, no. 11, pp. 2248–2254, 2003.

[26] C. Michael and M. Ismail, "Statistical modeling of device mismatch for analog MOS integrated circuits," *IEEE Journal of Solid-State Circuits*, vol. 27, no. 2, pp. 154–166, 1992.

[27] M. Conti, P. Crippa, S. Orcioni, and C. Turchetti, "Statistical modeling of MOS transistor mismatch based on the parameters' autocorrelation function," *Proc. 1999 IEEE International Symposium on Circuits and Systems*, vol. 6, pp. 222–225, 1999.

[28] Q. Zhang, J. Liou, J. McMacken, J. Thomson, and P. Layman, "Modeling of mismatch effect in submicron MOSFETs based on BSIM3 model and parametric tests," *IEEE Electron Device Letters*, vol. 22, no. 3, pp. 133–135, 2001.

[29] Q. Zhang, J. Liou, J. McMacken, J. Thomson, and P. Layman, "SPICE modeling and quick estimation of MOSFET mismatch based on BSIM3 model and parametric tests," *IEEE Journal of Solid-State Circuits*, vol. 36, no. 10, pp. 1592–1595, 2001.

[30] P. Drennan and C. McAndrew, "Understanding MOSFET mismatch for analog design," *Proc. of the 2002 Bipolar/BiCMOS Circuits and Technology Meeting*, pp. 449–452, 2002.

[31] J. Bastos, M. Steyaert, A. Pergoot, and W. Sansen, "Mismatch characterization of submicron MOS transistors," *Analog Integrated Circuits and Signal Processing*, vol. 12, no. 2, pp. 95–106, 1997.

[32] J.-B. Shyu, G. Temes, and F. Krummenacher, "Random error effects in matched MOS capacitors and current sources," *IEEE Journal of Solid-State Circuits*, vol. SC-19, no. 6, pp. 948–955, 1984.

[33] C. Abel, C. Michael, M. Ismail, C. Teng, and R. Lahri, "Characterization of transistor mismatch for statistical CAD of submicron CMOS analog circuits," *Proc. 1993 IEEE International Symposium on Circuits and Systems*, pp. 1401–1404, 1993.

[34] S.-C. Wong, J.-K. Ting, and S.-L. Hsu, "Characterization and modeling of MOS mismatch in analog CMOS technology," *Proc. of the 1995 International Conference on Microelectronic Test Structures*, pp. 171–176, 1995.

[35] S.-C. Wong, K.-H. Pan, D.-J. Ma, M. Liang, and P. Tseng, "On matching properties and process factors for submicrometer CMOS," *Proc. of the 1996 International Conference on Microelectronic Test Structures*, pp. 43–47, 1996.

[36] S.-C. Wong, K.-H. Pan, and D.-J. Ma, "A CMOS mismatch model and scaling effects," *IEEE Electron Device Letters*, vol. 18, no. 6, pp. 261–263, 1997.

[37] T. Serrano-Gotarredona and B. Linares-Barranco, "Mismatch characterization of submicron MOS transistors," *Analog Integrated Circuits and Signal Processing*, vol. 21, no. 3, pp. 271–296, 1999.

[38] T. Serrano-Gotarredona and B. Linares-Barranco, "A new 5-parameter MOS transistors mismatch model," *Proc. 6th IEEE International Conference on Electronics, Circuits and Systems*, pp. 315–318, 1999.

[39] T. Serrano-Gotarredona and B. Linares-Barranco, "A methodology for MOS transistor mismatch parameter extraction and mismatch simulation," *Proc. 2000 IEEE International Symposium on Circuits and Systems*, pp. 109–112, 2000.

[40] T. Serrano-Gotarredona and B. Linares-Barranco, "A new strong inversion 5-parameter transistor mismatch model," *Proc. 2000 IEEE International Symposium on Circuits and Systems*, pp. 381–384, 2000.

[41] M.-F. Lan and R. Geiger, "Modeling of random channel parameter variations in MOS transistors," *Proc. 2001 IEEE International Symposium on Circuits and Systems*, vol. 1, pp. 85–88, 2001.

[42] *HP4063 manual*. Agilent Technologies.

[43] H. Tuinhout, "Design of matching test structures," *Proc. of the 1994 International Conference on Microelectronic Test Structures*, pp. 21–27, 1994.

[44] A. Steegen, M. Stucchi, A. Lauwers, and K. Maex, "Silicide induced pattern density and orientation dependent transconductance in MOS transistors," *International Electron Devices Meeting 1999*, pp. 497–500, 1999.

[45] S. Chetlur, S. Sen, E. Harris, H. Vaidya, I. Kizilyalli, R. Gregor, and B. Harding, "Influence of passivation anneal position on metal coverage dependent mismatch and hot carrier reliability," *Proc. of 7th International Symposium on Physical and Failure Analysis of Integrated Circuits*, pp. 21–24, 1999.

[46] H. Tuinhout and M. Vertregt, "Characterization of systematic MOSFET current factor mismatch caused by metal CMP dummy structures," *IEEE Transactions on Semiconductor Manufacturing*, vol. 14, no. 4, pp. 302–310, 2001.

[47] R. Gregor, "On the relationship between topography and transistor matching in an analog CMOS technology," *IEEE Transactions on Electron Devices*, vol. 39, no. 2, pp. 275–282, 1992.

[48] G. Badenes, C. Perelló, A. Rupp, E. Vandamme, E. Augendre, S. Pochet, and L. Deferm, "Optimisation of critical parameters in a low cost high performance deep submicron CMOS technology," *Proc. of the 29th European Solid-State Device Research Conference*, pp. 628–631, 1999.

[49] F. Forti and M. Wright, "Measurement of MOS current mismatch in the weak inversion region," *IEEE Journal of Solid-State Circuits*, vol. 29, no. 2, pp. 138–142, 1994.

[50] M. Denison, A. Pergoot, and M. Tack, "Prediction of MOS matching in weak and moderate inversion from threshold matching in strong inversion," *Proc. of the 28th European Solid-State Device Research Conference*, pp. 648–651, 1998.

[51] T. Mizuno, "Influence of statistical spatial-nonuniformity of dopant atoms on threshold voltage in a system of many MOSFETs," *Japanese Journal of Applied Physics, Part 1*, vol. 35, no. 2B, pp. 842–848, 1996.

[52] S. Sun and J. Plummer, "Electron mobility in inversion and accumulation layers on thermally oxidized silicon surfaces," *IEEE Journal of Solid-State Circuits*, vol. SC-15, no. 4, pp. 562–573, 1980.

[53] S. Schwarz and S. Russek, "Semi-empirical equations for electron velocity in silicon: part I-bulk," *IEEE Transactions on Electron Devices*, vol. ED-30, no. 12, pp. 1629–1633, 1983.

[54] S. Schwarz and S. Russek, "Semi-empirical equations for electron velocity in silicon: part II-MOS inversion layer," *IEEE Transactions on Electron Devices*, vol. ED-30, no. 12, pp. 1634–1639, 1983.

[55] S. Takagi, A. Toriumi, M. Iwase, and H. Tango, "On the universality of inversion layer mobility in Si MOSFET's: part I-effects of substrate impurity concentration," *IEEE Transactions on Electron Devices*, vol. 41, no. 12, pp. 2357–2362, 1994.

[56] S. Takagi, A. Toriumi, M. Iwase, and H. Tango, "On the universality of inversion layer mobility in Si MOSFET's: part I-effects of surface orientation," *IEEE Transactions on Electron Devices*, vol. 41, no. 12, pp. 2363–2368, 1994.

[57] J. Hauser, "Extraction of experimental mobility data for MOS devices," *IEEE Transactions on Electron Devices*, vol. 43, no. 11, pp. 1981–1988, 1996.

[58] R. van Langevelde and F. Klaassen, "Effect of gate-field dependent mobility degradation on distortion analysis in MOSFET's," *IEEE Transactions on Electron Devices*, vol. 44, no. 11, pp. 2044–2052, 1997.

[59] E. Li, C. Liu, and H. Ng, "Effect of statistical variation on threshold voltage in narrow-channel MOSFETs," *Electronic Letters*, vol. 26, no. 17, pp. 1390–1391, 1990.

[60] E. Felt, A. Narayan, and A. Sangiovanni-Vincentelli, "Measurement and modeling of MOS transistor current mismatch in analog IC's," *Proc. of 1994 IEEE International Conference on Computer Aided Design*, pp. 272–277, 1994.

[61] M. Steyaert, J. Bastos, R. Roovers, P. Kinget, W. Sansen, B. Graindourze, A. Pergoot, and E. Janssen, "Threshold voltage mismatch in short-channel MOS transistors," *Electronic Letters*, vol. 30, no. 18, pp. 1546–1548, 1994.

[62] C. Kühn, S. Marksteiner, T. Kopley, and W. Weber, "New method for verification of analytical device models using transistor parameter fluctuations," *International Electron Devices Meeting 1997*, pp. 145–148, 1997.

[63] S. Lovett, M. Welten, A. Mathewson, and B. Mason, "Optimizing MOS transistor mismatch," *IEEE Journal of Solid-State Circuits*, vol. 33, no. 1, pp. 147–150, 1998.

[64] K. Okada, H. Onodera, and K. Tamaru, "Layout dependent matching analysis of CMOS circuits," *Analog Integrated Circuits and Signal Processing*, vol. 25, no. 3, pp. 309–318, 2000.

[65] A. Maxim and M. Gheorghe, "A novel physical based model of deep-submicron CMOS transistors mismatch for monte carlo spice simulation," *Proc. 2001 IEEE International Symposium on Circuits and Systems*, vol. 5, pp. 511–514, 2001.

[66] J. Schmitz, H. Tuinhout, A. Montree, Y. Ponomarev, P. Stolk, and P. Woerlee, "Gate polysilicon optimization for deep-submicron MOSFETs," *Proc. of the 29th European Solid-State Device Research Conference*, pp. 156–159, 1999.

[67] A. Van den Bosch, *High resolution, high speed CMOS current-steering digital-to-analog converters*. K.U. Leuven, 2003.

[68] H. Thibieroz and A. Duvallet, "Mismatch characterization and modelization of deep submicron CMOS transistors," *Proceedings of the SPIE*, vol. 3881, pp. 121–128, 1999.

[69] A. Van den Bosch, M. Steyaert, and W. Sansen, "The extraction of transistor mismatch parameters: the CMOS current-steering D/A converter as a test structure," *Proc. 2000 IEEE International Symposium on Circuits and Systems*, pp. 745–748, 2000.

[70] U. Schaper, C. Linnenbank, and R. Thewes, "Precise characterization of long-distance mismatch of CMOS devices," *IEEE Transactions on Semiconductor Manufacturing*, vol. 14, no. 4, pp. 311–317, 2001.

[71] M. Hamer, "First-order parameter extraction on enhancement silicon MOS transistors," *IEE Proceedings I (Solid-State and Electron Devices)*, vol. 133, no. 2, pp. 49–54, 1986.

[72] C. Mourrain, B. Cretu, G. Ghibaudo, and P. Cottin, "New method for parameter extraction in deep sub-micrometer MOSFETs," *Proc. of the 2000 International Conference on Microelectronic Test Structures*, pp. 181–186, 2000.

[73] S. Kubicek, W. Henson, A. De Keersgieter, G. Badenes, P. Jansen, H. van Meer, D. Kerr, A. Naem, L. Deferm, and K. De Meyer, "Investigation of intrinsic transistor performance of advanced CMOS devices with 2.5nm NO gate oxides," *International Electron Devices Meeting 1999*, pp. 823–826, 1999.

[74] H. Tuinhout, "Characterisation of systematic MOSFET transconductance mismatch," *Proc. of the 2000 International Conference on Microelectronic Test Structures*, pp. 131–136, 2000.

[75] T. Cochet, T. Skotnicki, G. Ghibaudo, and A. Poncet, "Lateral dependence of dopant-number threshold voltage fluctuations in MOSFETs," *Proc. of the 29th European Solid-State Device Research Conference*, pp. 680–683, 1999.

[76] N. Sano, K. Matsuzawa, A. Hiroki, and N. Nakayama, "Influence of statistical spatial-nonuniformity of dopant atoms on threshold voltage in a system of many MOSFETs," *Japanese Journal of Applied Physics, Part 2*, vol. 41, no. 5B, pp. 552–554, 2002.

[77] A. Asenov, "Random dopant threshold voltage fluctuations in 50 nm epitaxial channel MOSFETs: A 3d 'atomistic' simulation study," *Proc. of the 28th European Solid-State Device Research Conference*, pp. 300–303, 1998.

[78] A. Asenov, A. Brown, J. Davies, and S. Saini, "Hierarchical approach to "atomistic" 3-D MOSFET simulation," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 18, no. 11, pp. 1558–1565, 1999.

[79] A. Asenov, G. Slavcheva, A. Brown, J. Davies, and S. Saini, "Quantum mechanical enhancement of the random dopant induced threshold voltage fluctuations and lowering in sub 0.1 micron MOSFETs," *International Electron Devices Meeting 1999*, pp. 535–538, 1999.

[80] A. Asenov and S. Saini, "Suppression of random dopant-induced threshold voltage fluctuations in sub-0.1-μm MOSFET's with epitaxial and δ-doped channels," *IEEE Transactions on Electron Devices*, vol. 46, no. 8, pp. 1718–1724, 1999.

[81] D. Frank, Y. Taur, M. Ieong, and H.-S. Wong, "Monte Carlo modeling of threshold variation due to dopant fluctuations," *Proc. of 1999 Symposium on VLSI Technology*, pp. 169–170, 1999.

[82] A. Asenov and S. Kaya, "Effect of oxide interface roughness on the threshold voltage fluctuations in decano MOSFETs with ultrathin gate oxides," *Proc. 2000 International Conference on Simulation of Semiconductor Processes and Devices*, pp. 135–138, 2000.

[83] A. Asenov, G. Slavcheva, A. Brown, J. Davies, and S. Saini, "Increase in the random dopant induced threshold fluctuations and lowering in sub-100 nm MOSFETs due to quantum effects: A 3-d density-gradient simulation study," *IEEE Transactions on Electron Devices*, vol. 48, no. 4, pp. 722–729, 2001.

[84] A. Asenov, G. Slavcheva, A. Brown, R. Balasubramaniam, and J. Davies, "Statistical 3d 'atomistic' simulation of decano MOSFETs," *Superlattices and Microstructures*, vol. 27, no. 2/3, pp. 215–227, 2000.

[85] C.-R. Ryou, S. Hwang, H. Shin, C.-H. Lee, Y. Park, and H. Min, "Three-dimensional simulation of discrete oxide charge effects in 0.1 μm MOSFETs," *Solid-State Electronics*, vol. 45, pp. 1165–1172, 2001.

[86] H. Yamamoto, Y. Okada, and N. Sano, "Quantitative prediction of threshold voltage fluctuations in sub-100 nm MOSFETs by a new dopant model," *Proc. of the Device Research Conference*, pp. 171–172, 2001.

[87] T. Ezaki, T. Ikezawa, A. Notsu, K. Tanaka, and M. Hane, "3D MOSFET simulation considering long-range Coulomb potential effects for analyzing statistical dopant-induced fluctuations associated with atomistic process simulator," *Proc. 2002 International Conference on Simulation of Semiconductor Processes and Devices*, pp. 91–94, 2002.

[88] T. Ezaki, T. Ikezawa, and M. Hane, "Investigation of realistic dopant fluctuation induced device characteristics variation for sub-100nm CMOS by using atomistic 3D process/device simulator," *International Electron Devices Meeting 2002*, pp. 311–314, 2002.

[89] W. Gross, D. Vasileska, and D. Ferry, "Three-dimensional simulations of ultrasmall metal-oxide-semiconductor field-effect transistors: The role of the discrete impurities on the device terminal characteristics," *Journal of Applied Physics*, vol. 91, no. 6, pp. 3737–3740, 2002.

[90] G. Iannaccone and E. Amirante, "Quantum and semiclassical modeling of the threshold voltage dispersion due to random dopants in deep submicron MOSFETs," *Proc. of the 2002 2nd IEEE Conference on Nanotechnology*, pp. 197–200, 2002.

[91] Y. Oda, Y. Ohkura, K. Suzuki, S. Ito, H. Amakawa, and K. Nishi, "Statistical fluctuation analysis by Monte Carlo ion implantation method," *Proc. 2002 International Conference on Simulation of Semiconductor Processes and Devices*, pp. 199–202, 2002.

[92] T. Ezaki, T. Ikezawa, A. Notsu, K. Tanaka, and M. Hane, "Three dimensional MOSFET simulation for analyzing statistical dopant-induced fluctuations associated with atomistic process simulator," *IEICE Transactions on Electronics*, vol. E86-C, no. 3, pp. 409–415, 2003.

[93] Y. Oda, Y. Ohkura, K. Suzuki, S. Ito, H. Amakawa, and K. Nishi, "Statistical threshold voltage analysis by Monte Carlo ion implantation method," *IEICE Transactions on Electronics*, vol. E86-C, no. 3, pp. 416–420, 2003.

[94] V.-Y. Thean, M. Sadd, and E. White Jr., "Effects of dopant granularity on superhalo-channel MOSFET's according to two- and three-dimensional computer simulations," *IEEE Transactions on Nanotechnology*, vol. 2, no. 2, pp. 97–101, 2003.

[95] Y. Yasuda, M. Takamiya, and T. Hiramoto, "Separation of effects of statistical impurity number fluctuations and position distribution on V_{th} fluctuations in scaled MOSFETs," *IEEE Transactions on Electron Devices*, vol. 47, no. 10, pp. 1838–1842, 2000.

[96] Y. Tsividis, *Operation and modeling of the MOS transistor.* McGraw-Hill international editions, 2nd ed., 1999.

[97] Y. Taur and T. Ning, *Fundamentals of modern VLSI devices.* Cambridge university press, 1998.

[98] N. Collaert, *Alternative transistor structures: Modeling and optimisation of the vertical advanced heterojunction MOSFET.* K.U. Leuven, 2000.

[99] J. Brews, "Carrier-density fluctuations and the IGFET mobility near threshold," *Journal of Applied Physics*, vol. 46, no. 5, pp. 2193–2203, 1975.

[100] T. Ando, A. Fowler, and F. Stern, "Electronic properties of two-dimensional systems," *Reviews of Modern Physics*, vol. 54, no. 2, pp. 437–672, 1982.

[101] M. Lundstrom, *Fundamentals of carrier transport*. Cambridge university press, 2nd ed., 2000.

[102] H. Shin, G. Yeric, A. Tasch, and C. Maziar, "Physically-based models for effective mobility and local-field mobility of electrons in MOS inversion layers," *Solid-State Circuits*, vol. 34, no. 6, pp. 545–552, 1991.

[103] K. Takeuchi, "Channel size dependence of dopant-induced threshold voltage fluctuation," *Proc. of 1998 Symposium on VLSI Technology*, pp. 72–73, 1998.

[104] *Medici manual*. Avant!

[105] Y. Yasuda, M. Takamiya, and T. Hiramoto, "Threshold voltage fluctuations induced by statistical 'position' and 'number' impurity fluctuations in bulk MOSFETs," *Superlattices and Microstructures*, vol. 28, no. 5/6, pp. 357–361, 2000.

[106] T. Tanaka, T. Usuki, Y. Momiyama, and T. Sugii, "Direct measurement of V_{th} fluctuations caused by impurity positioning," *Proc. of 2000 Symposium on VLSI Technology*, pp. 136–137, 2000.

[107] T. Tanaka, T. Usuki, T. Futatsugi, Y. Momiyama, and T. Sugii, "V_{th} fluctuations induced by statistical variation of pocket dopant profile," *International Electron Devices Meeting 2000*, pp. 271–274, 2000.

[108] A. Asenov, "Random dopant induced threshold voltage lowering and fluctuations in sub 50 nm MOSFET's: a statistical 3d 'atomistic' simulation study," *Nanotechnology*, vol. 10, pp. 153–158, 1999.

[109] *International technology roadmap for semiconductors*. Sematech, 2001 ed.

[110] A. Asenov and S. Saini, "Influence of the polysilicon gate on the random dopant induced threshold voltage fluctuations in sub 100 nm MOSFETs with thin gate oxides," *Proc. of the 29th European Solid-State Device Research Conference*, pp. 188–191, 1999.

[111] R. Difrenza, P. Llinares, G. Morin, E. Granger, and G. Ghibaudo, "A new model for threshold voltage mismatch based on the random fluctuations of dopant number in the MOS transistor gate," *Proc. of the 31th European Solid-State Device Research Conference*, pp. 299–302, 2001.

[112] A. Asenov and S. Saini, "Polysilicon gate enhancement of the random dopant induced threshold voltage fluctuations in sub-100 nm MOSFET's with ultrathin gate oxide," *IEEE Transactions on Electron Devices*, vol. 47, no. 4, pp. 805–812, 2000.

[113] R. Difrenza, J. Vildeuil, P. Llinares, and G. Ghibaudo, "Impact of grain number fluctuations in the MOS transistor gate on matching performance," *Proc. of*

the 2003 International Conference on Microelectronic Test Structures, pp. 244–249, 2003.

[114] J. Brews, "Theory of the carrier-density fluctuations in an IGFET near threshold," *Journal of Applied Physics*, vol. 46, no. 5, pp. 2181–2192, 1975.

[115] A. Asenov, S. Kaya, and J. Davies, "Oxide thickness variation induced threshold voltage fluctuations in decano MOSFETs: a 3D density gradient simulation study," *Superlattices and Microstructures*, vol. 28, no. 5/6, pp. 507–515, 2000.

[116] G. Slavcheva, J. Davies, A. Brown, and A. Asenov, "Potential fluctuations in metal-oxide-semiconductor field-effect transistors generated by random impurities in the depletion layer," *Journal of Applied Physics*, vol. 91, no. 7, pp. 4326–4334, 2002.

[117] J. Zhao, H. Chen, and C. Teng, "Investigation of charging damage induced V_t mismatch for submicron mixed-signal technology," *Proc. of the Reliability Physics Symposium*, pp. 33–36, 1996.

[118] J. Bastos, M. Steyaert, R. Roovers, P. Kinget, W. Sansen, B. Graindourze, A. Pergoot, and E. Janssens, "Mismatch characterization of small size MOS transistors," *Proc. of the 1995 International Conference on Microelectronic Test Structures*, pp. 271–276, 1995.

[119] A. V. d. Bosch, M. Steyaert, and W. Sansen, "A high density matched hexagonal transistor structure in standard CMOS technology for high speed applications," *Proc. of the 1999 International Conference on Microelectronic Test Structures*, pp. 212–215, 1999.

[120] J. Dubois, J. Knol, M. Bolt, H. Tuinhout, J. Schmitz, and P. Stolk, "Impact of source/drain implants on threshold voltage matching in deep sub-micron CMOS technologies," *Proc. of the 32nd European Solid-State Device Research Conference*, pp. 115–118, 2002.

[121] J.-S. Goo, Q. Xiang, Y. Takamura, F. Arasnia, E. Paton, P. Besser, J. Pan, and M.-R. Lin, "Band offset induced threshold variation in strained-Si nMOSFETs," *IEEE Electron Device Letters*, vol. 24, no. 9, pp. 568–570, 2003.

[122] K. Takeuchi, R. Koh, and T. Mogami, "A study of the threshold voltage variation for ultra-small bulk and SOI CMOS," *IEEE Transactions on Electron Devices*, vol. 48, no. 9, pp. 1995–2001, 2001.

[123] S. Xiong and J. Bokor, "Sensitivity of double-gate and FinFET devices to process variations," *IEEE Transactions on Electron Devices*, vol. 50, no. 11, pp. 2255–2261, 2003.

[124] D. He, H. Solak, W. Li, and F. Cerrina, "Extreme ultraviolet and x-ray resist: Comparison study," *Journal of Vacuum Science & Technology B*, vol. 17, no. 6, pp. 3379–3383, 1999.

[125] S. G.M., M. Stewart, V. Singh, and C. Willson, "Spatial distribution of reaction products in positive tone chemically amplified resists," *Journal of Vacuum Science & Technology B*, vol. 20, no. 1, pp. 185–190, 2002.

[126] P. Oldiges, Q. Lin, K. Petrillo, M. Sanches, M. Ieong, and M. Hargrove, "Modeling line edge roughness effects in sub 100 nanometer gate length devices," *Proc. 2000 International Conference on Simulation of Semiconductor Processes and Devices*, pp. 131–134, 2000.

[127] T. Linton, M. Chandhok, B. Rice, and G. Schrom, "Determination of the line edge roughness specification for 34 nm devices," *International Electron Devices Meeting 2002*, pp. 303–306, 2002.

[128] S. Kaya, A. Brown, A. Asenov, D. Magot, and T. Linton, "Analysis of statistical fluctuations due to line edge roughness in sub-0.1μm MOSFETs," *Proc. 2001 International Conference on Simulation of Semiconductor Processes and Devices*, pp. 78–81, 2001.

[129] A. Asenov, S. Kaya, and A. Brown, "Intrinsic parameter fluctuations in decananometer MOSFETs introduced by gate line edge roughness," *IEEE Transactions on Electron Devices*, vol. 50, no. 5, pp. 1254–1260, 2003.

[130] C. Diaz, H.-J. Tao, Y.-C. Ku, A. Yen, and K. Young, "An experimentally validated analytical model for gate line-edge roughness (LER) effects on technology scaling," *IEEE Electron Device Letters*, vol. 22, no. 6, pp. 287–289, 2001.

About the Authors

Jeroen A. Croon was born in Delft, The Netherlands, on March 18, 1975. In 1998 he received a Master's degree in applied physics from the Delft University of Technology. His thesis involved the optimization of the receiving coil and pre-amplifier of an Overhauser MRI system at 77 K and at room temperature. After this he started working in IMEC, Belgium, and in 2004 he obtained a Ph.D. degree in electrical engineering from the Catholic University of Leuven, also in Belgium, for his study of the matching properties of deep submicron MOSFETs. His current research in IMEC involves the characterization of MOSFETs fabricated on germanium substrates, and the study of variability in advanced CMOS processes. In 2005 he will start working in the compact modeling group of Philips Research in Eindhoven, The Netherlands.

Willy Sansen has received the MSc degree in Electrical Engineering from the Katholieke Universiteit Leuven in 1967 and the PhD degree in Electronics from the University of California, Berkeley in 1972.
In 1972 he was appointed by the National Fund of Scientific Research (Belgium) at the ESAT laboratory of the K.U.Leuven, where he has been a full professor since 1980. During the period 1984-1990 he was the head of the Electrical Engineering Department. Since 1984 he has headed the ESAT-MICAS laboratory on analog design, which counts about sixty members and which is mainly active in research projects with industry. He is a fellow of the IEEE and is a member of several boards of directors.
In 1978 he was a visiting professor at Stanford University, in 1981 at the EPFL Lausanne, in 1985 at the University of Pennsylvania, Philadelphia, in 1994 at the T.H. Ulm and in 2004 at Infineon, Villach. Prof.Sansen is a member of several editorial and program committees of journals and conferences. He is cofounder and organizer of the

workshops on Advances in Analog Circuit Design in Europe. He is a member of the executive and program committees of the IEEE ISSCC conference. He was program chair of the ISSCC-2002 conference.
He has been involved in design automation and in numerous analog integrated circuit designs for telecommunications, consumer electronics, medical applications and sensors. He has been supervisor of over fifty PhD theses in these fields. He has authored and coauthored twelve books and more than 550 papers in international journals and conference proceedings.

Herman E. Maes was born in Leuven, Belgium on August 15, 1947. He received the M.Sc. degree in electrical engineering in 1971 and the Ph.D. degree in 1974, both from the Katholieke Universiteit Leuven in Belgium. From 1971 until 1974, he was a Research Assistant (Fellow of the National Fund of Scientific Research of Belgium, NFWO) in the laboratory for Physics and Electronics of the University of Leuven. In 1974, he was granted a CRB fellowship by the Belgian American Educational Foundation and spent 14 months at the Electrical Engineering Research Laboratory of the University of Illinois, Urbana, as a Research Associate. From 1975 until 1985, he was with the ESAT laboratory at the University of Leuven as a Senior Research associate of the Belgian National Fund for Scientific Research and a lecturer at the University. Since 1985, he has been a Professor at the University of Leuven. In 1985, he joined the newly established R&D Laboratory of the Interuniversity Micro-Electronics Center (IMEC) in Leuven, Belgium, as Head of Analysis and Reliability. In 1990 he became an Associate Vice-President in IMEC and in 1998 Vice-President, heading the Division on 'Silicon Technology and Device Integration'. Since 2002 he is heading the Division on "Industrialization and Training in Micro-electronics" at IMEC. He has been elected a Fellow of the IEEE (Institute of Electrical and Electronics Engineers) in 1997 *for contributions in the field of non-volatile silicon memory devices and for contributions to MOS Reliability Physics.* H.E. Maes has authored or coauthored more than 380 international technical papers including 8 book chapters and more than 390 conference papers including more than 40 invited papers. He has been guiding 30 students to the Ph.D. degree over the past 15 years.