MASTER

Simulation and optimization of a 0.25 mum CMOS process

Souren, B.J.C.

Award date:
1995

Link to publication
Simulation and Optimization
of a 0.25\textmu m CMOS process

B.J.C. Souren

Concerns:
Graduation Report
Faculty of Electrical Engineering
Eindhoven University of Technology
Department Elektronische Bouwstenen (EEA)

Place:
Philips Natuurkundig Laboratorium, Eindhoven

Period of Work:
June 1994-January 1995

Report Number:
EEA510

Coach:
Dr. M.J. van Dort

Graduation Coach:
Prof. Dr. ir. F.M. Klaassen

Philips Natlab.

TUE

The department of Electrical Engineering of the Eindhoven University of Technology is not responsible for the contents of graduation reports.
Abstract

Simulators are intensively used in optimizing new generation CMOS processes. At the Natuurkundig Laboratorium of Philips Research the sector Silicon Technology has started a project to produce a 0.25µm device. There are two reasons for looking to this specific process and the evaluation of the software: 1) CMOS devices are continuously under development. 2) The simulation software is developing and new versions with new options are introduced. In order to automate the simulations of the 0.25µm process new software is used. It is important to know whether new optimization techniques can reduce the simulation time. With the above simulations the optimum setting of the process parameters has to be extracted.

The aim of the report is to investigate how the new experimental software can speed up the optimization of the CMOS process by using design methods based on statistics. Another point of investigation concerns the result of the simulations: What is the best possible 0.25µm CMOS process setting?

The new software programs NORMAN and DEBORA automate the CMOS simulation process. It works fast and it reduces the time needed before a model is extracted. DEBORA uses a model to automatically optimize a process with respect to performance and reproducibility. The ability to do simulations in NORMAN is incorporated, but the transformed variables cannot yet be used in DEBORA. This needs to be changed. Another negative point of the software is that, although it works automatically, much handling on files is needed, which makes it user unfriendly. A great deal of expertise of individual simulation programs is needed to work with NORMAN and DEBORA.

The statistical method RSM is a good method to make a procedure describing the input and output parameters. It is also a method to investigate the resulting model and to get accurate results. The design methods reduce the necessary amount of simulations drastically. There are many possible designs, but a great expertise is needed to use them. They would better be used by just selecting the right design by the right kind of model.

The result of the optimization of the 0.25µm CMOS process is a setting of the input parameters where the responses are in a defined output range. This range is larger than the given specifications of the responses. The output range is enlarged because otherwise no useful point in the input range can be found. The resulting optimum is found by using the target criterium. In this optimum a sensitivity analysis is done, which shows the process sensitivity and the order of the input parameters. The input setting can be used as starting parameters for the real process.

**Keywords:**
- CMOS
- SUB MICRON
- RSM
- TOD
- SENSITIVITY
Table of contents

1 Introduction 9
   1.1 Motivation 9
   1.2 The aim of the report 10
   1.3 How this report is set up 10

2 The real and simulated process 11
   2.1 Introduction 11
   2.2 Description of the 0.25μm CMOS process 11
   2.3 Boundary conditions of device design 12
   2.4 The software programs 14
   2.5 The doping profile 15
   2.6 The device simulation via MINIMOS 17
   2.7 Programs which use design techniques to reduce the sensitivity of the process 20
      2.7.1 Introduction 20
      2.7.2 NORMAN 20
      2.7.3 DEBORA 21

3 The design methods 23
   3.1 Introduction 23
   3.2 The Response Surface Method 23
   3.3 How to use the response surface? 25
   3.4 The statistical methods for making a design 26
   3.5 The design methods 31
   3.6 How to do simulations 34

4 The way the simulations are performed 36
   4.1 Introduction 36
   4.2 Determine the factors to be varied 36
   4.3 The simulations of the n-channel device 37
      4.3.1 The first design 37
      4.3.2 Design with a shift of input 40
      4.3.3 The design with a larger input area 41
      4.3.4 The second order model 42
   4.4 The simulations of the p-channel device 43
   4.5 Conclusion 44

5 Results of the optimized process 46
   5.1 Introduction 46
   5.2 The way to optimize the CMOS-process 46
   5.3 Which optimum can be found? 47
   5.4 Sensitivity of the responses 50
      5.4.1 Total sensitivity in two points 50
      5.4.2 Sensitivity per input factor 54
Introduction

5.5 The optimum for the real transistor 59
   5.5.1 Accuracy of the model 59
   5.5.2 Accuracy of the real process 59
   5.5.3 The simulated optimum used in the real process 59
5.6 Conclusion 60

6 Conclusion 61

References 62

Appendix 1: Simulation of the doping profile 63
Appendix 2: Two-level fractional factorial design for \( k \) variables and \( N \) runs[9] 64
Appendix 3: NORMAN input file 65
1. Introduction

1.1 Motivation

Simulators are intensively used in optimizing new generation CMOS processes. At the Natuurkundig Laboratorium of Philips Research the sector Silicon Technology has started a project to produce a 0.25\(\mu\)m device. Simulations are used to speed up the development of the process and to gain physical insight into the process. There are two reasons for looking to this specific process and the evaluation of the software:

1 CMOS devices are continuously under development.
There are three important reasons for making the devices smaller:
- With the improving knowledge and control of the process steps it is possible to reduce the size.
- The smaller size can give better performance.
- The 0.25\(\mu\)m size is a new target for the minimum feature.

2 The simulation software is developing and new versions with new options are being introduced.
The following four options are important developments:
- A new trend is the automated optimization of CMOS processes with respect to performance and reproducibility [1]. This is also called the design for manufacturability, DFM [2].
- Especially process sensitivity is an important aspect in optimization [3]. Automated optimization is possible by using software which couples various simulation programs.
- Special design techniques which are based on statistical methods can improve the effectiveness of the optimization program[4].
- Experimental software, called NORMAN and DEBORA, has become available[12]. These programs use a statistical design method like the Response Surface Method, RSM[9].

In order to automate the simulations of the 0.25\(\mu\)m process new software is used. It is important to know whether the new optimization techniques can reduce the simulation time. With these simulations the optimum setting of the process parameters has to be extracted.
1.2 The aim of the report

The aim of this report is to investigate how the new available experimental software can speed up the optimization of the CMOS process by using design methods based on statistics. An other point of investigation is the result of the simulations: What is the best possible 0.25μm CMOS process setting?

The following 3 steps are used to give answers to the questions:

• The new software is used to do all simulations automatically.
• The investigation is performed by studying the design techniques and the corresponding statistical methods.
• To optimize the process the results of the simulations are compared with the specifications.

1.3 How this report is set up

Chapter 2 gives an overview of the real CMOS process and how it is simulated. The statistical methods used for the new design techniques are explained in chapter 3. Chapter 4 explains how the design techniques are used to select the simulations and how this leads to an optimum process. The results of the simulations and the relation between the process and the output parameters is described in chapter 5. Conclusions on how the software should be used to quickly optimize the process and how this process looks like is given in chapter 6.
2 The real and simulated process

2.1 Introduction

The process flow of the 0.25μm CMOS process needs to be optimized to satisfy the specifications for the device parameters. A good estimate of the process parameter settings for the real experiments is achieved by simulating the complete device. The complete simulation consists of three steps:

- the process simulation which generates a doping profile
- the device modelling to find the electric characteristics
- the device parameter extraction

These steps have to be repeated for different process parameter settings. A new program called NORMAN automates these steps. The results of the simulations is a list of device parameter settings under different process conditions. These results are used by another program called DEBORA to make an empirical model between the input and output parameters. The model is optimized to find the optimum setting of the process flow.

The aim of this chapter is:

- to discuss which CMOS process is used and how it works
- to discuss the software used to optimize the process to satisfy the specifications

This chapter gives an overview of the 0.25μm CMOS process which needs to be optimized with the new available software. In addition to the process and the specifications, the different software programs and their interactions are described.

The variations of the simulated process are limited to the channel doping profile under the gate. The device behaviour is tested in two dimensions. The software programs NORMAN and DEBORA are still in an experimental stage.

2.2 Description of the 0.25μm CMOS process

On a p-type substrate thick LOCOS is grown selectively. The transistor areas are defined between the LOCOS, see Fig. 2.1b. First a sacrificial oxide, SACOX, is grown to protect the Si-surface from the implantations and to have a clean surface for growing the gate oxide after the SACOX is removed. A Twin Retrograde Well technology is used for the well formation. Boron implantations are used for the n-channel device. A retrograde p-well is formed by the first high energy implantation. The implant provides a low sheet resistance, while maintaining fairly low surface doping. The low resistance reduces the latch up sensitivity. The Channel Stopper implant (CS) suppresses possible inversion layers underneath the LOCOS isolation between
transistors. The third implantation for the n-channel device is the Threshold Voltage Implantation. It is a low energy Boron implantation.

The p-channel well is implanted with Phosphorous. Similar to the three equivalent implantations which are used in the n-channel device, the p-channel device has an additional implantation, the Anti Punch Through implantation. The Threshold Voltage Implantation for the p-channel is not a Phosphorous but a low energy Arsenic implantation. Compared to Phosphorous, Arsenic has the advantage not to diffuse so much and to create a shallow dope at the surface of the Silicium.

After the implantations, the SACOX is removed and the gate oxide is thermally grown. The poly silicon layer is deposited and the subsequent etching forms the gate. After deposition of a TEOS layer which restores the gate integrity, which has been deteriorated during the gate etch, the drain and source are implanted.

### 2.3 Boundary conditions of device design

CMOS must be designed to meet the electrical specification given by the 0.25μm project, see Tab. 2.1. These device parameters are taken to be parameters of the MOS MODEL 9 and they determine the performance of the transistor. The quality increases as the demands on the performance are achieved. The threshold voltage $V_T$, the saturation current $I_{SAT}$ and $I_{OFF}$ are the most important parameters.

The power supply voltage of the 0.25μm device is 2.5V. This low voltage is chosen in order to reduce the power dissipation of the transistor. In fact device scaling influences
Tab. 2.1  Device parameter specification

<table>
<thead>
<tr>
<th>Nominal length of the device is: $L_{EFF}=0.25 \mu m$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Power Supply Voltage</strong></td>
</tr>
<tr>
<td><strong>Gate Oxide</strong></td>
</tr>
<tr>
<td>$V_{T} N10/10$</td>
</tr>
<tr>
<td>$V_{T} N_{nom}$</td>
</tr>
<tr>
<td>$V_{T} P10/10$</td>
</tr>
<tr>
<td>$V_{T} P_{nom}$</td>
</tr>
<tr>
<td>$\Delta L_n$ per side</td>
</tr>
<tr>
<td>$\Delta L_p$ per side</td>
</tr>
<tr>
<td>$I_{SAT,n}$</td>
</tr>
<tr>
<td>$I_{SAT,p}$</td>
</tr>
<tr>
<td>$I_{d,OFF}$ ($V_D=2.75V$, $L_{EFF}=0.2\mu m$)</td>
</tr>
<tr>
<td>Body-effect coefficient</td>
</tr>
<tr>
<td>Subthreshold slope</td>
</tr>
</tbody>
</table>

the supply voltage[5]. Because the electric field region near the drain is extremely small heating of the carriers becomes important. The power supply voltage needs to be reduced to achieve a long device lifetime.

The process is designed with a nominal gate oxide thickness of 6 nm. This oxide thickness is well controllable.

The threshold voltages, have to be specified for the long channel device and for the short channel device. The specification for $V_{T}$ of the short channel device is a minimum voltage. The long channel $V_{T}$ must meet the target value. Ideally the short channel $V_{T}$ should be the same as the long channel $V_{T}$, but short channel effects lower its value.

The maximum gate drain overlap $\Delta L_{n,p}$, or gate source overlap is specified to be 50nm.

The current drive $I_{SAT,n}$, of the device is specified as the minimum current per unit width at maximum gate and drain voltage. Because the electron mobility is higher than the hole mobility, the current of the $n$-channel device is higher than the $p$-channel device.

The leakage current $I_{OFF}$, is defined as the maximum current per unit width under worst case conditions. The drain voltage is 10% increased and the effective channel length is 20% decreased. This corresponds to a worst case situation for the transistor.
The body-effect coefficient $K$, must have a maximum value of 0.3 $V^{-1}$.

The sub threshold slope is defined as the change in gate bias for which the current changes by one decade in the sub threshold region. This will be called $V_{SWING}$ and must be as small as possible.

In the procedure of parameter extraction the transistor parameters $\theta_A$ and $\beta$ are automatically extracted. There is no specification, but they can be monitored. The mobility reduction parameter $\theta_A$ should be small and the gain factor $\beta$ should be large.

### 2.4 The software programs

In Fig. 2.2 a block diagram of the integrated chain of process simulator, device simulator and parameter extraction is given.

![Block diagram of simulators interaction](image)

**Fig. 2.2 Interaction of simulators with NORMAN and DEBORA**

The program NORMAN takes control of the in and outputs. The simulators perform different tasks. The flowchart data is read by the process simulator SUPREM3. It generates a one dimensional doping profile under the gate. The device structure is completed by adding a two dimensional doping file to it via the program SUPMIN. The resulting two dimensional doping file is used by the device simulator MINIMOS to calculate the electrical characteristics. PAREX fits this data to MOS MODEL9 [6] and extracts the compact model parameters. After the simulations, with varied process
parameters, the data is used by DEBORA to make an empirical model. This model is used by DEBORA to optimize the process.

2.5 The doping profile

The SUPREM3 input file is the description of the process flow of the real process. The most important process steps which affect the channel profile are incorporated. These process steps are explained here.

The one dimensional SUPREM3 simulation program is the best choice in order to generate a doping profile. The simulated profile is visualized by the line AA' in the crosscut of Fig. 2.1.d. The rest of the doping profile, containing the area of the source and drain, is added later to get a 2-dimensional doping profile. This extra data is generated with the 2-dimensional simulator SUPREM4. This simulator could also be used to directly generate the complete profile, but the disadvantage is the long simulation time of about 45 minutes per simulation. Because of the expected large amount of simulations to get an accurate model between the process and the electrical parameters the CPU time should be reduced as much as possible. The advantage of SUPREM3 is the speed. One minute simulation time to calculate the channel profile is normal. The disadvantage is that it describes the transistor only in one dimension.

The process simulator SUPREM3 can simulate each process step in one dimension:
• It is used to simulate the channel profile underneath the gate.
• The implantation steps and all furnace anneals in the subsequent processing influence the doping underneath the gate.
• The forming of the LOCOS does not influence the channel profile because all implantations are done after the formation of the field oxide. However the implantation of the Channel Stopper must be adjusted to have it's maximum doping concentration just underneath the LOCOS.

The process steps in the SUPREM3 input file

In the CMOS process the n- and p-channel transistors have many common process steps, but the process steps which determine the channel doping under the gate are performed with other ions and therefore with other implantation parameters. The p-channel transistor is doped with Phosphorus and Arsenic and the n-channel transistor with Boron. The implantation steps are described by two parameters: the energy and the dose of the implanted ion. An example of the one dimensional doping profile is given in appendix 1. The SUPREM3 file for the n-channel and the p-channel transistor is given in the NORMAN input file in appendix 6.
There are four implantations:

- First the retrograde well is formed by an implantation step with a high energy, see Tab. 2.2.
- For both transistors the thickness of the LOCOS determines the Channel Stop implantation. The maximum implantation dose should be just underneath the LOCOS and the implantation dose should be adjusted so that the threshold voltage of the parasitic field transistor is too high for conduction. Therefore the implantation of the Channel Stopper is not variable. Since other implantation ions are used, the energy and dope concentrations is different for the $n$- and $p$-channel transistor.
- The APT implantation for the $p$-channel transistor should be high enough to suppress punch through, but an increase of the dope underneath the channel influences the threshold voltage or saturation current; this dependence should be as low as possible.
- The Threshold voltage implantation has a low energy level to keep the implantation shallow.

### Tab. 2.2 The process parameters and their initial setting

<table>
<thead>
<tr>
<th>Process parameters</th>
<th>$n$-channel</th>
<th>$p$-channel</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>dose [cm$^2$]</td>
<td>energy [keV]</td>
</tr>
<tr>
<td>1 Well</td>
<td>$1.0 \times 10^{13}$</td>
<td>180</td>
</tr>
<tr>
<td>2 CS</td>
<td>$7.0 \times 10^{12}$</td>
<td>90</td>
</tr>
<tr>
<td>3 APT</td>
<td>$1.5 \times 10^{12}$</td>
<td>100</td>
</tr>
<tr>
<td>4 $V_T$</td>
<td>$5 \times 10^{12}$</td>
<td>16</td>
</tr>
<tr>
<td>5 Channel length</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 Oxide thickness</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In Tab. 2.2 the difference between the $n$-channel and $p$-channel process parameter setting is listed. The common process steps are the definition of the LOCOS and the growth of the gate oxide. The thickness of the field oxide determines the Channel Stop implantation in the $n$-channel and $p$-channel transistor. The implantation must have its maximum dose underneath the LOCOS to stop a parasitic channel underneath the LOCOS.

Because different ions for implantation are used the implantation parameters are different for both channels. In fact the well implantation, the Anti Punch Through implantation and the $V_T$ implantation are different, see Tab. 2.2.

After the implantations, a furnace step is done to activate the dope and to remove the
damage of the implantations. This causes a diffusion of the implanted ions in both transistors. This part of the process is important for the simulations, because it changes the doping profile of the channel.

The CMOS process has a low temperature budget, meaning a low value for the product of temperature and time. There are two reasons for this:

• To make a thin gate oxide the oxide has to be grown slowly with a lower temperature.
• To make shallow Source and Drain junctions the diffusion of the dope during the furnace anneal needs to be minimized.

2.6 The device simulation via MINIMOS

MINIMOS is the MOS device simulator. MINIMOS uses the two dimensional doping profile together with a defined channel length and width of the transistor to calculate the electric characteristics of the MOSFET.

The electron and hole concentration, the lateral and transversal electric field and the current and the potential are determined for a given bias. The bulk, source, drain and the gate bias have to be specified.

Correction for $L_{\text{EFF}}$

Since the specifications listed in Tab. 2.1 are defined for this length, it is important to simulate the $I$-$V$ characteristics for an effective channel length of exactly 0.25μm. To determine the effective gate length the Terada Method is used. In Fig. 2.3 this method is visualized.

Since the dope concentration in the Source and Drain is high, a variation in the total resistance is caused by the channel length. By calculating the total resistance for different gate potentials as a function of the gate length the gate drain overlap can be determined. The poly gate length in Fig. 2.3 is equal to the gate drain plus gate source overlap if the total resistance is constant for different gate biases, so $L_{\text{POLY}}=2\Delta L$. With the resulting gate drain overlap $\Delta L$, the poly gate length $L_{\text{POLY}}$ can be determined for every effective gate length $L_{\text{EFF}}$.

The extracted parameters

The specifications of the transistor are determined by the parameters given in Tab. 2.1. These parameters are calculated with MINIMOS and the method to do this is described here.

To plot the electrical characteristic function $I_D$-$V_{GS}$, several points have to be calculated of this characteristic: The Drain current $I_D$ is calculated for different back biases as a function of the gate voltage $V_{GS}$. An example is given in Fig. 2.4. The calculated points are fitted to a model function by the program PAREX, see Fig. 2.2.
The real and simulated process

Fig. 2.3 The Terada method used to determine $L_{\text{EFF}}$

Fig. 2.4 $I_D-V_{GS}$ characteristic
From Fig. 2.4, PAREX extracts five important compact MOS model parameters:

- the threshold voltage $V_T$ is determined by extrapolating the function to zero current
- the gain is determined for $V_D = 0.1$ from the slope of the function (for $V_D$ the current equation can be simplified to only first order effects of $(V_{GS} - V_T)$)
- $\theta_A$ and $\beta$ are determined by fitting the MINIMOS points to the simplified $I_T V_{gs}$ relation, for $V_{DS} < 1$, $V_{GS} > V_T$:

\[
I_{DS} = \beta \cdot G \cdot \frac{(V_{GS} - V_T) \cdot V_{DS}}{1 + \theta_A (V_{GS} - V_T)}
\]  

(2.1)

where $\theta_A$ and $\beta$ can be considered as process-dependent parameters.
- the Body effect coefficient $K$

In order to obtain reliable simulation results for the 0.25μm process the MINIMOS mobility model was carefully calibrated using a 0.35μm process for which Silicon is already available[7].

The threshold voltage $V_{th}$ determined by MINIMOS is normally too low, compared to the real process. The offset is 50mV[8]. Another offset is caused by the high surface dope concentration. A high concentration at the surface influences the energy bands and this results in a lower calculated threshold voltage[8]. This difference is determined once before the actual simulations by comparing the threshold voltage of one process, with and without the quantum correction. The resulting offset is a mean correction value but the resulting threshold voltage is more accurate.

The saturation current $I_{SAT}$ is determined for a bias $V_D = V_{DD}$ and $V_G = V_{DD}$.

The current $I_{OFF}$ is the current determined for $L_{EFF} = 0.2μm$, $V_G = 0V$ and $V_{DD} = 2.75V$. Because of strong dependence of $I_{OFF}$ on $V_{DD}$, this current is strongly non linear in the process parameters. This makes it difficult to model $I_{OFF}$ with a first or second order model. Instead of calculating the current at $V_G = 0$, the gate voltage is determined for the specified maximum current. This voltage is determined by MINIMOS using the option V-THRESH. In this case the current is $I_{D,OFF} = 10pA/μm$. The minimum current given by the specifications in Tab. 2.1.

In the subthreshold region the gate voltage swing is determined for a change of the current for one decade. Between $I_{D1} = 1 \times 10^{-8}W/L$ and $I_{D2} = 1 \times 10^{-9}W/L$. For this range of the drain current the relation between $I_D$ and $V_{GS}$ is exponential.
2.7 Programs which use design techniques to reduce the sensitivity of the process

2.7.1 Introduction

In the Technology CAD (TCAD) environment two programs are very important:
- NORMAN automates the simulation of the process flow and the other steps to extract device parameters.
- DEBORA first uses the simulations to calculate an empirical model between the process parameters and the device parameters. The model is further used to optimize the process sensitivity.

In this paragraph these two programs are described. First an input file for NORMAN is described and then the use of DEBORA is described.

2.7.2 NORMAN

Input
The input file of NORMAN consists of three parts:
- definition of the input process parameters
- description of the simulation path
- description of the simulator input files

In the first part the variable names and their range, which are used in the whole NORMAN file, are declared. These names are variables which are used in for example the SUPREM3 process simulator for the implantation energy.

The range of the input factors is made discrete, by a low and high value. A design file is written by NORMAN which describes the parameter setting per simulation. In the design file the selected variables are varied according to a specified design method. These design methods are treated in chapter 3.

In the second part of the NORMAN file the order of simulations is defined. The simulation sequence to follow is given in Fig. 2.2. The simulators are run by NORMAN, which handles the in and output of the programs. Eventually the last step is the collection of all simulation results. Norman collects for all output parameters the simulation results for the simulations specified in the design file.

The third part of the NORMAN file describes the input files of the individual simulators. Here the variable names, described in the first part, are used. The NORMAN input file for the n-channel and the p-channel is given in appendix 6.

Output
The output of NORMAN consists of the design file plus one output file per response. The design file describes all parameter settings per simulation. The output file for each response gives the calculated responses for the simulations in the design file.
Method
NORMAN uses design methods described in chapter 3 to reduce the amount of simulations needed to build a model. The program is used to extract the device parameters for many different process parameter settings. Because an accurate model is needed to relate the input and output parameters, the amount of simulations depends strongly on the amount of parameters. The more input parameters the more simulations are needed to make a correct model. Design methods like the Response Surface Method (RSM) and Target Oriented Design (TOD) minimize the amount of simulations needed to get accurate models.

2.7.3 DEBORA

Input
The input file consist of four parts:
• the order of the model used
• the specifications of the output parameters
• the description of the response files made by NORMAN
• the method used by DEBORA to find an optimum or perform a sensitivity analyses

Output
The output depends on the method used by DEBORA. If an optimum is searched the result is given in output files describing the input and output settings. For a sensitivity analysis the output is given by a file consisting of the calculated outputs of the response.

Method
The best process setting is found by optimizing the model to reduce the sensitivity of the device parameters for variations of the process parameters. The aim for Design For Manufacturability (DFM), is twofold: The device parameters must satisfy the specification and the sensitivity to process fluctuations should be as small as possible.

Find an optimum
DEBORA is a program which optimizes different response functions with the same input parameters. The response functions are modeled for a number of input variables, based on a set of simulated data and a design file. The simulated date and the design file is made by NORMAN.

The model is a first or second order Taylor series expansion. Also reduced Taylor series can be made by fitting the simulation to a function that is related to the first order and some second order input parameters.

There are two optimization criteria for the specified points:
• the target value of the response function
• the sensitivity of the response function for input parameters

The first criteria, the target, is incorporated in DEBORA by defining a target deviation function. Each response function is given a weight of importance. The sum of the squared deviation between the responses in the specific point and the target
values is a selection criterium which should be minimal. So the best point is found where the sum of the deviation between the weighted response functions and the target value is minimal.

The second criteria, the sensitivity of the response function, is incorporated by defining a sensitivity function. The directive of each response function for each input parameter is determined for the specified design points. The response function is given a weight of importance. Also here the point with a minimal sum of the squared weighted directives is the best design point.

These target and sensitivity criteria can be used separately but can also be combined to find the best design point. This is achieved by giving the first and second criteria a weight of importance.

Sensitivity analysis
The design points can be used to make a sensitivity analysis. The input parameters can be given a standard deviation, which is known from physical consideration or is based on experiments. If \( x \) points were calculated in a 3-dimensional space these points would be in a sphere around the design point. The total deviation of the response resulting from the input deviation gives an indication of the sensitivity of the particular response, which is important to determine the manufactureability of the process. The sensitivity of the simulated process parameters for the 0.25\( \mu \)m process are given in Tab. 2.3. With DEBORA these results are easily obtained. DEBORA calculates the output value for a given number of simulations within the deviation of the input factors. From these outputs the mean and the standard deviation is calculated.

<table>
<thead>
<tr>
<th>process parameter</th>
<th>setting</th>
<th>max deviation</th>
<th>unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Implantation energy</td>
<td>energy &lt; 50</td>
<td>1%</td>
<td>keV</td>
</tr>
<tr>
<td></td>
<td>energy &gt; 50</td>
<td>2%</td>
<td>keV</td>
</tr>
<tr>
<td>Implantation dose</td>
<td>( 1 \times 10^{12} )</td>
<td>3%</td>
<td>cm(^2)</td>
</tr>
<tr>
<td>channel length</td>
<td>0.25</td>
<td>4%</td>
<td>( \mu )m</td>
</tr>
<tr>
<td>oxide thickness</td>
<td>6</td>
<td>1%</td>
<td>nm</td>
</tr>
</tbody>
</table>
3 The design methods

3.1 Introduction

A design describes the $n$-input parameter settings per simulation and one setting represents a point in a $n$-dimensional space. The simplest design is made by varying one input factor at a time, but this method is deceiving if it is used to determine the effect on the output [9]. The theory of the Response Surface Method (RSM) and Target Oriented Design (TOD) are promising techniques which indicate which simulations and how many should be performed to determine a correct model between the input and output parameters. This chapter explains these methods and shows how they are used. Statistical methods are used to evaluated them.

Experiments of a specific process are performed for two reasons:
- to learn about a process and to get better control of it
- to get a relation between the process input variables and the output variables

The purpose of this chapter is to show which method should be used to reduce the amount of simulations needed to relate the input and output parameters in a model. First the relation between the input and output parameters is described. Based on literature the statistical basic principles are explained, from which the design methods follow. The design methods are explained and then the exploration steps of an empirical model are described.

3.2 The Response Surface Method

The Response Surface Method includes statistical techniques for empirical model building and model exploitation [9]. By careful design and analysis of experiments it seeks to relate a response, or output variable to the levels of a number of input variables which affect it.

Important for the Response Surface Method is:
- the definition of a response surface
- the model used to describe the response
- the error of the model
- the amount of experiments needed

The definition of a response surface
The response function describes the relation between the input and output parameters. If there is no knowledge about the physical relationship this function can be approximated by an empirical relation.

For more than two variables the response is called a response surface, see for example Fig. 3.1. The two inputs are called factors. In Fig. 3.1 factor one is the
The design methods

pressure and factor two is the temperature. As the factors change in magnitude the response function changes.

\[ y = f(\xi) + \epsilon \]  

\[ \xi \] is an input vector with \( k \) elements \((\xi_1, \xi_2, \ldots, \xi_k)\). The function \( f(\xi) = \eta \) is the mean response at a particular condition \((\xi_1, \xi_2, \ldots, \xi_k)\). An actual observed response result, \( y \), falls in a statistically distribution around its mean value \( \eta \). The expected value of \( y \) equals \( \eta \), that is, \( E(y) = \eta \). The nature of this expectation function is unknown from physical consideration and is therefore replaced by the graduating function \( g(\xi) \), which is a polynomial of a certain degree.

It is useful to normalize the input value \( \xi \). If the input \( \xi_i \) is defined as \( \xi_{i0} \pm S_i \), where \( \xi_{i0} \) is the centre of the region then the normalized variable is

\[ x_i = \frac{\xi_i - \xi_{i0}}{S_i} \]  

The response function of a polynomial of order 2 with \( k \) inputs and coefficients \( \beta \) is:

\[ g(x) = \beta_0 + (\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k) + 
(\beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \ldots + \beta_{k-1,k} x_{k-1} x_k) + (\beta_{11} x_1^2 + \ldots + \beta_{kk} x_k^2) \]  

The number of coefficients \( \beta \) for a polynomial of order 2 with \( k \) inputs is \( \frac{1}{2} (k+1)(k+2) \) and increases rapidly as the number of input variables increases. The
determination of the coefficients is achieved by fitting the function to the simulated or experimented points.

The error of the model
There are two types of errors:

- The experimental error $\varepsilon$ is the difference $y-\eta$ and consists of three errors:
  - measurement error
  - observational error
  - random error from basic variability in the experimental material
- The systematic error $E(y)-g(\xi)$ arising from the inability of the graduating function $g(\xi)$ to exactly match the expectation function $E(y)=f(\xi)$.

In simulations the experimental error $\varepsilon$ is zero. However the systematic error is important in simulations.

The amount of experiments needed
Two important conclusions for the amount of experiments are the following:

1. the higher the degree of the approximating function, the more closely the Taylor series can approximate the true function but the more experiments are needed.
2. the smaller the region $R$ over which the approximation needs to be made, the better is the approximation possible with a polynomial function of given degree.

3.3 How to use the response surface?

The response is described as function of an input vector $x$, see (3,3). In the $n$-dimensional space a combination of $x$ variables has to be found which satisfies the specification given by the output $y_i$.

The response surfaces can be used in three ways:

1. To show how a particular response surface $y_i$ is affected by a set of variables $x$ over some specific region of interest.
2. To discover what settings of $x$ will give a result which simultaneously satisfies the specifications for a number of responses $y_1, y_2, ..., y_m$.
3. To explore the space of the $x$ variables to find the maximum response of $y_i$ and to determine the nature of this maximum.

Since the response surface is determined for a specific input range, the second and third point are likely to need a movement away from the initial experimental region before the objective is obtained.

It is also important to determine the local geography of the region under interest. The exploration of such ridges can provide alternative near-optimal processes and thus allow the near optimization of more than one response.

The models use to describe the relation between the input and output parameters are
local approximations of the responses via polynomials. The Response Surface Method learns the investigator more about:

- the amount of replications needed to achieve sufficient precision in case the simulations are performed as real experiments
- the location of the experimental region of most interest
- the appropriate scaling and transformations for the input and output variables
- the degree of complexity of an approximating function and hence of the designs needed at various stages

These response surfaces can afterwards also be helpful to determine a physical model of the response. This is a true functional form of the response rather than an approximation by a graduating function.

3.4 The statistical methods for making a design

A design is already explained to be a collection of input parameter settings. These input parameters are called factors. The relation between the input factors and the response is determined by experiments or simulations. It is important that the factor setting of each simulation is carefully chosen, so that the coefficients $\beta$ of (3.3) are correct.

This paragraph gives the basic information of the statistical methods needed to make a design according to the RSM. These designs are explained in the next paragraph. The methods can also be used in the program NORMAN.

The factorial design
The statistical methods to reduce the number of experiments is based on the theory of factorial design. A factorial design is a class of experimental design. The most interesting inputs of a specific process are chosen to be factors which influence the response function most. A design is obtained in $k$ factors by choosing $n_1$ levels of factor 1, $n_2$ levels of factor 2, and $n_k$ levels of factor $k$ and then selecting the $n=n_1n_2\ldots n_k$ runs obtained by tracking all possible combinations of the levels selected. Fig. 3.2 shows a 3·2·2-factorial design in the factors Temperature, Energy, Dose. The temperature is varied over three levels while the dose and energy are varied over two levels. In this design all points in the 3-dimensional space are factor settings for this specific design.
The main effect
At the exploratory stage of an investigation the two-level-factorial design is important. In Tab. 3.1 the design of a 2-level-3-factorial design, a $2^3$-design, is given in standard order.

The standard order is obtained by writing

- alternate - and + signs in the column headed $x_1$
- alternate pairs --, ++ in the $x_2$ column
- alternate fours ----,++++ in the $x_3$ column

The -, represents the low and the +, the high factor value.

This $2^3$-design would give an almost identical figure as Fig. 3.2. All possible combinations of factors are chosen in the design. In table Tab. 3.1 each factor is changed from low to high four times. See for example factor 1. For all combinations of factor 2 an 3(four combinations), the factor 1 is varied from low to high. The effect of factor 1, which is defined as the difference between the output $y$ for the high and low factor value, can be calculated four times. The average difference in the level of response as one moves from the low to the high level of that variable is called the main effect of $x_1$. The linear contrast is denoted by $l_1$, and is an estimate of the effect of $x_1$ in this case. For a full factorial design the linear contrast equals the main or interaction effects. The main effect can be calculated from Tab. 3.1 and the results are given in Tab. 3.2.
The design methods

### Tab. 3.1 A $2^3$ factorial design

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_{12}$</th>
<th>$x_{13}$</th>
<th>$x_{23}$</th>
<th>$x_{123}$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>2,5</td>
</tr>
<tr>
<td>2</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>2,7</td>
</tr>
<tr>
<td>3</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>2,1</td>
</tr>
<tr>
<td>4</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2,4</td>
</tr>
<tr>
<td>5</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3,5</td>
</tr>
<tr>
<td>6</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>4,2</td>
</tr>
<tr>
<td>7</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3,1</td>
</tr>
<tr>
<td>8</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>3,9</td>
</tr>
</tbody>
</table>

The 2-level-3-factorial design is organized in standard order. The output $y$ is determined for the setting of column $x_1$, $x_2$, $x_3$. 'I' is the identity. The main effect of $x_3$ is calculated by summing the output $y$ for a '+' sign and subtracting the output $y$ for a '-' sign in the corresponding column. The sum is divided by the divisor, see Table 3.2.

### Tab. 3.2 Determine the contrast for the factors

<table>
<thead>
<tr>
<th>main effects</th>
<th>$l_1$</th>
<th>$l_2$</th>
<th>$l_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$l_1$</td>
<td>1/4[+(2,7+2,4+4,2+3,9)-(2,5+2,1+3,5+3,1)] = 0,5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$l_2$</td>
<td>2 = -0,35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$l_3$</td>
<td>3 = 1,25</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2-factor-interaction effects

| $l_{12}$ | 12 = 0,05 |
| $l_{13}$ | 13 = 0,25 |
| $l_{23}$ | 23 = 0,00 |

3-factor-interaction effect

| $l_{123}$ | 123 = 0,00 |

The effects are determined from Table 3.1.
Interaction effects

Two variables, \(x_1\) and \(x_3\), interact, if the effect of \(x_1\) changes at the two different levels of \(x_3\). This means that the effect of \(x_1\) depends on the setting of \(x_3\). The main effect was already calculated for all four settings of the other two factors. This can also be done for the interaction effect of \(x_{13}\). This is easily measured if an extra column \(x_{13}\) is made by multiplying the elements of column \(x_1\) and \(x_3\). The contrast \(l_{13}\), which is an estimate of the two factor effect \(x_{13}\), is then calculated, see Tab. 3.2. Also the three factor effect of \(x_{123}\), can be calculated in a similar way.

Response function

From the results of the main effects, the interaction effects and the mean value \(y\), the response function can be determined:

\[ y = 3.05 + 0.25x_1 - 0.17x_2 + 0.63x_3 \]  

(3.4)

The coefficients in this formula are half the factor effect of Tab. 3.2, because an effect was defined as the difference in response on moving from the -1 level to the +1 level of a given variable \(x_i\), which corresponds to the change in \(y\) if \(x_i\) is changed by two units. The constant is the mean response value of \(y\).

Blocking and fractionating factorial Designs

The reason for blocking the design is that not all experiments can be performed under exact the same conditions. An example could be a furnace step for a Silicon wafer. If there was only room for half the amount of wafers of the \(2^3\) factorial design in the furnace and the other half would be processed later under the same conditions, the effect of the furnace could change the output. Because the aim of the experiments is to relate the output \(y\) and the input factors \(x_1, x_2, x_3\), and not the relation between the output and the furnace, the low and high values of these factors should be mixed about the different furnace steps. So, looking at Tab. 3.1, it is not desired to put all experiments with a positive \(x_1\) in one block.

The column of \(x_{123}\) in Tab. 3.1 can be used to define two blocks in a \(2^3\) factorial design. From Tab. 3.2 can be seen that the \(x_{123}\) interaction effect is zero. Normally high order interactions are very low. The effect of the furnace step can be seen as a fourth factor \(x_4\). If the settings of factor \(x_4\) is the same as factor \(x_{123}\) then these factors are confounded. For high values of factor \(x_4\) the wafer is put in the first furnace step and for low values the wafer is put in the last furnace step.

Fractionate design

The defining equation for a blocked \(2^6\) factorial design can be, for example:

\[ I = 123456, \]

which is called the generator. The identity \(I\) is a column with all plus signs. The sign of the six factors in one block will multiply to give +1. The right part of the equation is the \(x_{123456}\) interaction column, as in Tab. 3.1. From this defining equation the interactions of the factors with other factors can be retrieved. For example the first factor is confounded with the \(x_{23456}\) effect. This is obtained by multiplying the
The design methods

generator with the appropriate factor

\[ 1 = 1 \cdot I = 1 \cdot 123456 = 23456 \]

and the \( x_{123} \) effect is confounded with the \( x_{456} \) interaction effect. Now the linear contrasts \( l_1 \), is not only an estimate of the factor effect of \( x_1 \), but also the factor effect \( x_{23456} \). Also the contrast of \( x_{123} \), which is \( l_{123} \), is an estimate of the effect of \( x_{123} \) and the effect of \( x_{456} \). This is written as:

\[
\begin{align*}
  l_1 & \rightarrow 1 + 23456 \\
  l_{123} & \rightarrow 123 + 456 
\end{align*}
\]
Resolution of a $2^{k-p}$ factorial design

The resolution $R$ of a $2^{k-p}$ fractional factorial design is the length of the shortest word in the defining equation. A $2^{12}_6$ design with defining relation

$I = 1234 = 3456 = 1256$

therefore is of resolution IV.

For a design of resolution $R$ accounts: no effects of order ($i < R$), are aliased with effects of orders smaller than $(R-i)$.

Given the generator above, corresponding to a resolution IV, the following contrasts are true:

\[
\begin{align*}
I_1 &\rightarrow 1 + 234 + 256 \\
I_2 &\rightarrow 2 + 134 + 156 \\
I_{12} &\rightarrow 12 + 56 \\
I_{234} &\rightarrow 234 + 1 + 256 + 13456
\end{align*}
\]

The first order effect are not aliased with other first order effects nor with second order effects. The $I_{12}$ interaction effect and other second order effects are aliased with other second order effects.

As a consequence, to estimate the coefficients of a first order design with no interactions of the main effects, a factor of at least resolution III should be used. In fact a resolution V should be used to get coefficients free of second order aliasing.

3.5 The design methods

The design is the collection of factor settings according to which the simulations are performed. With help of the statistical methods these designs are easily defined. There are designs for first order and second order models. In the program NORMAN these methods are incorporated.

Compared with the real experiments the time for doing simulations is drastically reduced, but is still important. With a standard process numerous input factors can be varied which give rise to many simulations and much simulation time. For 10 input factors with two levels already $2^{10} = 1024$ simulations have to be done. The corresponding simulation time is minimal a day if the simulations is limited to a one dimension model, which is too long. The process parameters are numerous and so a strong selection has to be made in choosing the most important factors.

The design which are treated in this paragraph are:

- Full Factorial 2 level
- Fractional Factorial 2 level
- Central Composite Faced, CCF
- Central Composite Inscribed, CCI
- Central Composite Circumscribed, CCC
The design methods

- Taguchi
- Target Oriented Design, TOD

After explaining these designs Tab. 3.3 shows the amount of simulations needed per method for a different amount of input variables.

**Full factorial 2 level and fractional factorial 2 level design**
These designs are already explained in the former paragraph. A complete collection of fractional 2 level designs is given in appendix 2. For 3 to 11 number of variables and for 4 to 128 number of simulations, a possible defining equation is given.

For example a $2_{10-6}^3$ fractional factorial design for 2 levels of 10 factors with resolution III only needs $2^k=16$ simulations. With these simulations the main effects can be calculated which are not aliased with the two factor interactions. For this case the confounding sentence in NORMAN could be:

```
EXP=FRA CONF=1235;2346;1347;1248;12349;12(10)
```

Here all defining equations are put in one confounding sentence.

**The central composite design**
A central composite design consists of a two-level factorial augmented with further points. In general, a composite design contains:

1. A 'cube', consisting of a $2^k$ factorial, or a $2_{-P}^{k-p}$ fractional factorial design, of resolution $R \geq 5$, replicated $r_c (\geq 1)$ times. There are thus $n_c = r_c 2^{k-p}$ such points.
2. A 'star', which is, $2k$ points $(\pm \alpha, 0, \ldots, 0), (0, \pm \alpha, 0, \ldots, 0), \ldots, (0, \ldots, 0, \pm \alpha)$ on the variable axes, replicated $r_s$ times, so that there are $n_s = 2kr_s$ points.
3. Centre points $(0, 0, \ldots, 0), n_o$ in number.

Replications can be necessary for real experiments, but for the simulations one is enough. An example of a central composite design for 3 factors is given in Fig. 3.3. The parameter $\alpha$ depends on the method used:

- $\alpha < 1 \quad \alpha = 0.42 \quad$ CCI = Central Composite Inscribed design
- $\alpha = 1 \quad$ CCF = Central Composite Faced design
- $\alpha > 1 \quad \alpha = 2.38 \quad$ CCC = Central Composite Circumscribed design
The choice between these three designs depends on the choice of the parameter $\alpha$:
- The central composite design can be used for fitting a second order model.
- Because the CCx design consists of a $2^k$ factorial design it is easy to use the central composite design as an extension of it. The $2^k$ design is than the basis.
- The centre point is a handy point, because it can be used for the 'curvature check', to check the accuracy of the model, fitted to the simulations.

**Taguchi method**
The Tagushi designs are special fractional factorial designs. A 6 factorial design is made by first generating a $2^3$ factorial design with the first 3 factors. The factors 4, 5 and 6 are aliased with the interaction effects of the $2^3$ factorial design. The design is still orthogonal. These designs are only used for purely linear models [10]. Since normally the design is used to investigate a response for the input factors to eventually build a 2nd order model, probably the Taguchi method is not necessary. So after these simulations are done, new simulations have to be done for the 2nd order.

**Target Oriented Design (TOD)**
The TOD is a procedure to reduce the amount of simulations needed [11]. Based on a preliminary limited set of simulations, the TOD procedure generates a design of new simulations.

A factorial design can be used as a preliminary design. If extra simulations are needed to explore the model and to find responses fit to the specifications the TOD can help.

Using the result from the first simulations, the new experimental TOD is calculated within the parameters space, so that the new simulations are expected to give outputs which are within specified constraints.
Tab. 3.3 shows the amount of simulations that are needed per method. The amount of simulations needed to calculate a second order model are 66 if 10 variables are used. The CCF design requires to simulate 1045 points to fit a model. If the TOD method is used, exactly 66 simulations can be done to generate a second order model. This is less accurate than compared to the CCF design, but is satisfactory because fewer simulations are chosen where the output region is expected to realize the specifications.

### 3.6 How to do simulations

The experimental designs are described and the model is defined. The best method needs to be chosen to perform the simulations and to get a satisfying model between the input and output. This paragraph gives a method to perform the simulations by using the design methods given. These design methods can be used by NORMAN.

The main question is how to use the design and how the results are obtained. The method of doing experiments is given by the RSM and other sources.

The simulations need to be performed to relate the in and output parameters in the empirical model described by (3,3).

The input parameters are chosen on the basis of process knowledge. The amount of parameters can be large if it is not clear which parameters are important. The range of the parameters should be given. This can be large but needs to be somewhere in the
area of the best value. The output parameters are limited by the specifications which need to be given, otherwise a reasonable area has to be specified.

The first experimental design is chosen on the basis of a expected linear model. For a pure linear model a factorial design of Resolution II is satisfactory. This kind of design has main effects that are not confounded with interaction effects. The interaction effects can be confounded.

Depending on the amount of input variables the table in appendix 2 can be used to chose a fractional factorial design. For 10 input factors a minimum of 16 simulations is described for fitting a first order model with a design of Resolution III. NORMAN ads one centre point to the design. A point at the middle of all factors.

These simulations are used to fit the first order model for all responses. Before looking to the desired range for the input parameters the accuracy of the model is tested with the $R^2$ parameters. If the accuracy is not good the input factors and the responses have to be checked:

- If the output value changes, over decades then there is need for transformation of the parameter. This is only possible in NORMAN for Central Composite designs, so for this preliminary research the output has to be transformed by a sub program or other trick, such as the $I_{OFF}$, $V_{PUNCH}$ transformation in paragraph 2.6.

- To increase the accuracy the input factor region can be decreased. The fit is then an estimate of a smaller output area. The decreased input parameters change the range of the output parameters. So new simulations have to be checked whether they satisfy to all output constraints.

- Another method to increase the accuracy is to increase the order of the model. For a second order of $k$ inputs $\frac{1}{2}(k+1)(k+2)$ coefficient and simulations are needed. To increase the amount of simulations two methods can be used. either a fold over of the first experimental design or the TOD technique.

If the output is completely out of range the desired functional behaviour could be different and the input range has to be changed. This is not easy. The input range can be increased to include the desired region in the used region. But if the accuracy decreases too much nothing can be said about the correctness of the output.
4 The way the simulations are performed

4.1 Introduction

The simulations are performed for the 0.25µm CMOS process with the one dimensional simulator. The simulations in this chapter correspond to the simulated experiments as described in chapter 2. There is also a list of the input and output parameters. If a process needs to be optimized quicker it means in this contents that the amount of simulations is reduced. The designer of a transistor has to meet target specifications. These targets are parameters of the compact transistor model, a set of mathematical relations describing the transistor. These parameters are in general dependent on the geometry of the device and on the technological process steps. In order to determine their influence on the compact model parameters, the most important technological process steps and geometrical parameters are varied.

The aim of this chapter is to show how an accurate model between the input and the output process parameters is achieved and to show which factors are varied under what conditions. This chapter is a an example of the use of the RSM in an IC-production process. It is important that the simulations are chosen carefully because the amount can increase rapidly. Another important reason is the correctness that has to be achieved.

The accuracy between the in and output of the process is important to:

- use the relation to optimize the output parameters
- predict the output changes related to input changes

The main question in this chapter is which simulations are used to achieve a satisfying model describing the CMOS transistor. First an inventory is given of the available input and output parameters, their ranges and accuracy. The different simulation steps are described separately. For every simulation the description, expectation, aim, results and conclusion is given. Based on the conclusion new simulations are performed.

4.2 Determine the factors to be varied

The CMOS device is described by two factor types:

- the factors which include the technological process steps
- the geometry dependent factors

The process factors are included in the SUPREM3 file to generate the one dimensional doping profile, see Tab. 4.1. The geometry of the device is incorporated in the MINIMOS simulation.
4.3 The simulations of the $n$-channel device

The $n$-channel device is simulated separately from the $p$-channel device. In this paragraph the simulations are described which are used to determine the model parameters of the $n$-channel transistor. The sub paragraphs describe the steps in the simulation process to find the correct input range corresponding to a accurate model satisfying the output specifications.

4.3.1 The first design

*Description A*

The first simulations are performed by varying all input factors. For the $n$-channel 10 factors are varied. The input range is limited to a small area. The first design is a fractional factorial design of Resolution III. From appendix 2 the defining equation is determined, and here given in NORMAN format:

$$\text{CONF}=1235;2346;1347;1248;12349;12(10)$$

The amount of simulations is 17, including one centre point added by NORMAN.

In this simulation there is an APT implantation in the $n$-channel, which was not foreseen in the process flow, but was introduced to test the influence of such an implantation.
The simulations

Aim A
The aim of the simulations is to find the response function relating the input and output. The output should satisfy the specifications given in Tab. 2.1.

1. the first simulations are not expected to give already an accurate model, because the response behaviour is not yet known and therefore the amount of simulations needed to assure an accurate model
2. the responses will give some information about the process; it can be used to decide for transformation of variables
3. the model coefficients can give an indication of the order of the input factors

Results A
Tab. 4.2 shows the result of four simulation designs. The input factors are given a ranking by calculating their importance for the responses. This is the mean effect per input for all responses. Also the accuracy of the response models is given in Tab. 4.2. The corresponding Least Square parameter $R^2$ has been explained in chapter 3. This parameter should be one, to assure the best accuracy. The response models of Design A are bad. Only the long channel $V_r$, $I_{SAT}$ and Swing are accurate. The correct order of the input factors is therefore doubtful.

<table>
<thead>
<tr>
<th>Input Factor</th>
<th>Importance of factors per design</th>
<th>Response</th>
<th>Response Accuracy $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>vtnedo</td>
<td>1.33</td>
<td>1.33</td>
<td>1.28</td>
</tr>
<tr>
<td>aptneco</td>
<td>2.28</td>
<td>2.27</td>
<td>2.21</td>
</tr>
<tr>
<td>dox</td>
<td>6.10</td>
<td>6.11</td>
<td>3.15</td>
</tr>
<tr>
<td>vtnccen</td>
<td>3.16</td>
<td>4.15</td>
<td>5.12</td>
</tr>
<tr>
<td>aptncen</td>
<td>4.15</td>
<td>3.16</td>
<td>4.14</td>
</tr>
<tr>
<td>csnccdo</td>
<td>5.12</td>
<td>5.12</td>
<td>6.10</td>
</tr>
<tr>
<td>csnccen</td>
<td>9.08</td>
<td>9.06</td>
<td>7.05</td>
</tr>
<tr>
<td>sacox</td>
<td>8.08</td>
<td>10.05</td>
<td>8.03</td>
</tr>
<tr>
<td>pwdo</td>
<td>10.06</td>
<td>8.06</td>
<td>9.03</td>
</tr>
<tr>
<td>pwen</td>
<td>7.09</td>
<td>7.07</td>
<td>10.02</td>
</tr>
</tbody>
</table>

*The term 1.33 means a first ranking for this factor, with a score of 0.33.

**$R^2$ is the Least Square parameter and this should ideally be one.

Conclusion A
The model is determined by doing 17 simulations. The model is not accurate enough
because the R² parameters are too small or even negative. The lack of accuracy can have several causes:

- the amount of simulations is too less
- the parameter range is too large
- the relation between the input and output is strongly nonlinear, meaning that interaction model terms are important

The order of the input factors can be determined from the model coefficients, but these coefficients are not accurate enough. New simulations have to be done to increase the accuracy of the models.

**Design B**

**Description B**

To increase the amount of simulations centre points are added, based on the TOD procedure. With this procedure 12 points are found which are added to the design to get a total of 29 simulations.

**Aim B**

The extra simulations should increase the accuracy of the model.

**Results B**

Tab. 4.2 shows the results. For all responses the response fit is improved, except for the long channel body coefficient $K$, the long channel $V_T$ and the saturation current $I_{SAT}$.

**Conclusion B**

By doing more simulations the accuracy of the model improves, but is not accurate for all parameters. More simulations are needed to improve the accuracy.

**Design C**

**Description C**

The amount of simulations is increased by doing another fractional factorial run of resolution III, see appendix 2. This is defined in NORMAN by:

```
CONF= -1235;2346;1347;1248;12349;12(10)
```

The first term is inverted compared to Design A, leading to another 16 simulations and a total of 45 simulations.

**Aim C**

The extra simulations should increase the accuracy of the model again.

**Result C**

The result of the extra simulations is diverse. Some response accuracy improves, like $I_{OFF}$ and the long channel $V_T$, but others get worse.
Conclusion C
Increasing the amount of simulations does not necessarily improve the accuracy of the response model. This can be explained for a model that may only be correct for a specific part of the input space. Also the behaviour of the response can be strongly nonlinear.
Extra simulation points can increase the accuracy, so an other TOD should be calculated, based on the 45 experiments of Design C.

Design D
Description D
Design C is increased with extra simulation points by using the TOD method. With this method 21 points are found. The total amount of simulations increases to 66.

Aim D
By doing extra simulations the model is expected to increase in accuracy.

Result D
Except for the long channel $\beta$ and the body coefficient $K$, the accuracy increases.

Conclusion D
Not all output parameters satisfy the constraints for the whole input range; Some are too low or too high. Therefore the input range has to be shifted to find a better setting, where all responses satisfy their constraints. Another option is to enlarge the input range so that all areas are included.

The process shows a strong variation in $I_{\text{app}}$ that means that an exponential relation between the input and output exists. A transformation is necessary to increase the accuracy of this response model.

New simulations have to be done to:
• increase the accuracy of the models
• satisfy the constraints by choosing the good input range

4.3.2 Design with a shift of input

Description
In order to achieve the correct output value corresponding to the specification, the input range of the parameters is shifted. The shift direction of the input factor is based on the results of DEBORA, the model from paragraph 4.3.1. Since DEBORA can not find any input setting that satisfies the constraints, the constraints are changed. The allowed output range is increased. With this increased output range, DEBORA finds input settings that satisfy these constraints. If the input setting of one factor reaches the boundary of the input range, the new input range is extended in that direction.
Aim
Simulations are necessary to determine the input range which satisfies to the constraints. There may be more than one shift necessary to achieve the correct input range.

Conclusion
Shifting the input range is not effective. The amount of output parameters is too big to find the correct input range to satisfy all constraints. An area needs to be found where all constraint satisfy the specification. If this area exist it can be found by simulating the complete possible area.

4.3.3 The design with a larger input area

To do simulations with a larger input range, this range has been determined in experiment A. Experiment B in this paragraph investigates the possibilities of this larger input range.

A Determination the range

Description
To simulate the transistor in the complete possible input range, this range has to be determined first. This can be done by investigating the two types of input factors: the process parameters and the geometry parameters. The process input parameters like the implantation dose and energy are investigated by checking their effect on the doping profile generated by the process simulator SUPREM3. When no effect can be seen the edge of the input range is reached. Geometry parameters like the thickness of the gate oxide have not been varied owing to the fact that the process specification is very stringent.

Aim
Expected is that the complete input range can be determined by viewing the doping profile generated by the process simulator.

Results
A range for the input parameters is found. Because the implanted doping concentration is determined by the implantation energy and the doping dose it is hard to determine both their minimum range. This is best explained with the $V_T$ implantation. For a too low value of the implantation energy, no increased doping concentration can be seen in the doping profile.

The APT implantation and the well implantation can be varied largely because they are implanted deeper than the $V_T$ implant. Only the CS implantation can not be varied much because it is related to the thickness of the LOCOS.
B Simulating with a larger input range

Description
The input range is increased according to the experiment A. To find any points in the input range where the responses satisfy the constraints, the output specification range is enlarged. The first design is a design of resolution $R=\text{III}$. Extra simulations are found by using a TOD method. Because the output range is enlarged the TOD method can find input settings, that satisfy the specifications. Extra simulations are used to make a more accurate model. DEBORA is used to find an optimum point that satisfies best to the enlarged constraints.

Aim
Expected is that the responses are accurate enough to investigate the input range on possible areas that satisfy the output constraints. Also expected is that the enlarged input range includes a point that satisfies all constraints of the larger output specification range.

Results
From the model follows that the dose and energy of the $V_T$ and APT implantation and the thickness of the oxide are the most important parameters. The optimum point found by using DEBORA with the enlarged output specification results in a setting for the less important parameters.

Conclusion
The optimum input setting that is found in the enlarged output specification can be used as a setting for next simulations.

4.3.4 The second order model

Description
The 5 most important electrical parameters and an extra geometrical parameter $L_{\text{EFF}}$ are varied to describe a model. A CCF design is used to generate the input settings.

Aim
The simulations should be enough to describe an accurate second order model.

Results
The results of this design are listed in Tab. 4.3. The input parameters with their range and the accuracy of the output parameters are given.
Tab. 4.3  CCF design for the n channel transistor

<table>
<thead>
<tr>
<th>n channel Factor</th>
<th>CCF-design range</th>
<th>unit</th>
<th>Response L=10µm, S=0,25µm</th>
<th>Response Accuracy $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>vtncdo</td>
<td>6e12-9e12</td>
<td>cm²</td>
<td>L</td>
<td>$\beta$</td>
</tr>
<tr>
<td>dox</td>
<td>5.8-6.2</td>
<td>nm</td>
<td>$K$</td>
<td>.98</td>
</tr>
<tr>
<td>varl</td>
<td>0.2-0.3</td>
<td>µm</td>
<td>$\theta_A$</td>
<td>.73</td>
</tr>
<tr>
<td>csenaptncdo</td>
<td>6e11-6e12</td>
<td>cm²</td>
<td>$V_T$</td>
<td>1</td>
</tr>
<tr>
<td>vtncen</td>
<td>5-50</td>
<td>keV</td>
<td>S</td>
<td>$\beta$</td>
</tr>
<tr>
<td>csenaptncen</td>
<td>60-90</td>
<td>keV</td>
<td>$I_{OFF}$</td>
<td>.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$V_{punch}$</td>
<td>1</td>
</tr>
<tr>
<td>sacox</td>
<td>10</td>
<td>nm</td>
<td>$I_{SAT}$</td>
<td>1</td>
</tr>
<tr>
<td>pwdo</td>
<td>1e13</td>
<td>cm²</td>
<td>$K$</td>
<td>.85</td>
</tr>
<tr>
<td>pwen</td>
<td>180</td>
<td>keV</td>
<td>$SWING$</td>
<td>.95</td>
</tr>
</tbody>
</table>

$R^2$ is the Least Square parameter and this should ideally be one.

Conclusion
The required accuracy of the model is achieved. This model can be used to optimize the process.

4.4 The simulations of the p-channel device

Similar to the n-channel the p-channel transistor is simulated and an input range is found that satisfies the enlarged output specifications. The model, determined from these simulations can be used to optimize the process. In Tab. 4.4 the resulting input range is given together with the accuracy of the models determined with a CCF design.
The simulations

4.5 Conclusion

The simulations have been performed for the CMOS process by separately doing simulations for the $n$-channel and the $p$-channel. Four steps can be distinguished:

- Based on a factorial design the simulations have been performed and a first order model has been made.
- The input range had to be changed to find the correct responses that satisfy the output constraints. This has been done by two methods:
  - By shifting the input range. This solves the problem for one response but an other response is deteriorated.
  - By increasing the input range. With a larger input range the complete range for the responses can be found. The input range corresponding to the optimum points can then be used as the starting points for improving the model.
- By doing simulations around the input range satisfying all constraints a more accurate model can be found by using less input parameters and a reduced input range.
- The last step is to increase the model by making a second order model. Therefore extra simulations have been performed. The output then satisfies to the larger specification range.

The resulting second order model is achieved with a limited small input range. The response models are accurate because enough simulations have been performed and
the model is of the second order. The restriction of the model is that the responses only satisfy the larger output specifications. The model can be used to optimize all responses to satisfy the specification. Probably not all responses can be within the small output specification. This is investigated in the next chapter.
5 Results of the optimized process

5.1 Introduction

The simulated process has been calculated in chapter 4. An accurate model is found that relates the input and output. The 0.25μm CMOS process needs to be optimized in order to realize the specifications given by Tab. 2.1 and to reduce the sensitivity of the output parameters to the process parameters.

In the optimization process two worlds can be distinguished. The simulated transistor and the real transistor. The first one has been optimized in the previous chapter. The difference between the real and the simulated transistor is minimized by taking MINIMOS parameters which are extracted from real experiments, see chapter 2.6. The target of this chapter is to analyze the results of the optimal process.

The main question of this chapter to show what is the optimal setting for the simulated process is and how this setting can be used for the real process. The best possible setting of the process with respect to the specifications and the sensitivities is discussed in this chapter.

1. The resulting device parameters are compared with the specifications. For some output parameters the correlation with the process parameters is given to show the trend.
2. To visualize the sensitivity of the process two process parameter settings are compared by showing the resulting device parameter spread. Only the most important device parameters are visualized to illustrate the optimization process.

5.2 The way to optimize the CMOS-process

To optimize the process there are two criteria for the output of the responses that have to be reduced:

1. the target deviation
2. the sensitivity to the input parameters

These criteria are used in the programme DEBORA as discussed in chapter 2.

To optimize the process there are several steps:

1. Find a model between in and output parameters; This has been done in chapter four with simulations.
2. Generate starting points that satisfy the output specifications. A starting point is a setting for the input factors.
3. Compare these points based on the two criteria as listed above.
If no starting points are found that satisfy all specifications, the output range must be enlarged. This can be done on the basis of physical consideration and knowledge of the process.

5.3 Which optimum can be found?

Because the specifications are to strict, no optimum can be found.

For the $p$-channel transistor the two most important parameters $V_T$ and $I_{SAT}$ conflict.

![Graphs showing $I_{sat}$ and $V_t$ as functions of implant dose and energy](image)

Fig. 5.1 $p$-channel, $L_{eff}=0.25\mu m$

According to Fig. 5.1 the most important input parameter of the $p$-channel transistor, the $V_T$-implantation dose should be low to realize the specification for $I_{SAT}>|-2mA|$ and the dose should be high for $V_T>|-0.5V|$.

With the above model not all specifications can be realized, but a parameter setting can be found for which the responses satisfy the enlarged specification. So they come close to an optimum. This optimum cannot be reached because of conflicting interests between the specifications for $V_T$ and $I_{SAT}$ as shown in Fig. 5.1.

Also the criterium for $V_{SWING}$ is critical. In Fig. 5.2 the $V_T$ implant dose should be low to keep $V_{SWING}$ low. This corresponds to the implantation dose of $I_{SAT}$ but conflicts with the $V_T$ requirement.
For the n-channel transistor the most important parameters $V_T$ and $I_{OFF}$ conflict. The most important input parameter (the $V_T$-implantation dose) should be high to realize the minimum specification for $I_{OFF}$. The dose should be lower to realize the exact long channel $V_T$. With this model not all specifications can be realized, but a parameter setting can be found for which the responses satisfy the enlarged specification. So they come close to an optimum. This optimum cannot be reached because of conflicting interests between the specifications $I_{OFF}$ and $V_T$. 

Fig 5.2 $V_{\text{swing}}$ variations
In Tab. 5.1 the results are listed for the p-channel transistor. Both the input range and the corresponding output setting are given. In Tab. 5.2 this is done for the n-channel transistor.

Tab. 5.1  The p-channel settings and the results

<table>
<thead>
<tr>
<th>Input Factor</th>
<th>p-channel</th>
<th>range</th>
<th>Optimum unit</th>
<th>$L=10\mu m, S=0.25\mu m$</th>
<th>Spec.</th>
<th>Optimum</th>
</tr>
</thead>
<tbody>
<tr>
<td>vtpedo</td>
<td></td>
<td>$1e12-3e12$</td>
<td>$2.36e12$</td>
<td>cm$^2$</td>
<td>L</td>
<td>$\beta$</td>
</tr>
<tr>
<td>dox</td>
<td></td>
<td>5.8-6.2</td>
<td>6</td>
<td>nm</td>
<td>$K$</td>
<td>$V^+$</td>
</tr>
<tr>
<td>varl</td>
<td></td>
<td>0.2-0.3</td>
<td>0</td>
<td>$\mu m$</td>
<td>$\theta_A$</td>
<td>$1/V$</td>
</tr>
<tr>
<td>apppedo</td>
<td></td>
<td>$2e11-3e12$</td>
<td>$2e11$</td>
<td>cm$^2$</td>
<td>$V_T$</td>
<td>$V$</td>
</tr>
<tr>
<td>vtpcen</td>
<td></td>
<td>20-60</td>
<td>42.82</td>
<td>keV</td>
<td>S</td>
<td>$\beta$</td>
</tr>
<tr>
<td>apppcen</td>
<td></td>
<td>$130-190$</td>
<td>190</td>
<td>keV</td>
<td>$V_{punch}$</td>
<td>$V$</td>
</tr>
<tr>
<td>sacox</td>
<td></td>
<td>10</td>
<td>10</td>
<td>nm</td>
<td>$I_{SET}$</td>
<td>mA</td>
</tr>
<tr>
<td>nwdno</td>
<td></td>
<td>$1e13$</td>
<td>$1e13$</td>
<td>cm$^2$</td>
<td>$K$</td>
<td>$V^+$</td>
</tr>
<tr>
<td>nwen</td>
<td></td>
<td>500</td>
<td>500</td>
<td>keV</td>
<td>$SWING$</td>
<td>$V$</td>
</tr>
</tbody>
</table>

$<1$ as low as possible, $>1$ as high as possible

Tab. 5.2  The n-channel results

<table>
<thead>
<tr>
<th>Input Factor</th>
<th>n-channel</th>
<th>range</th>
<th>Optimum unit</th>
<th>$L=10\mu m, S=0.25\mu m$</th>
<th>Spec.</th>
<th>Optimum</th>
</tr>
</thead>
<tbody>
<tr>
<td>vtnedo</td>
<td></td>
<td>$6e12-9e12$</td>
<td>$7.8e12$</td>
<td>cm$^2$</td>
<td>L</td>
<td>$\beta$</td>
</tr>
<tr>
<td>dox</td>
<td></td>
<td>5.8-6.2</td>
<td>6</td>
<td>nm</td>
<td>$K$</td>
<td>$V^+$</td>
</tr>
<tr>
<td>varl</td>
<td></td>
<td>0.2-0.3</td>
<td>0</td>
<td>$\mu m$</td>
<td>$\theta_A$</td>
<td>$1/V$</td>
</tr>
<tr>
<td>csenaltnedo</td>
<td></td>
<td>$6e11-6e12$</td>
<td>$6e12$</td>
<td>cm$^2$</td>
<td>$V_T$</td>
<td>$V$</td>
</tr>
<tr>
<td>vtnecn</td>
<td></td>
<td>5-50</td>
<td>35.6</td>
<td>keV</td>
<td>S</td>
<td>$\beta$</td>
</tr>
<tr>
<td>csenaltnecn</td>
<td></td>
<td>$60-90$</td>
<td>64</td>
<td>keV</td>
<td>$V_{punch}$</td>
<td>$V$</td>
</tr>
<tr>
<td>sacox</td>
<td></td>
<td>10</td>
<td>10</td>
<td>nm</td>
<td>$I_{SET}$</td>
<td>mA</td>
</tr>
<tr>
<td>pwo</td>
<td></td>
<td>$1e13$</td>
<td>$1e13$</td>
<td>cm$^2$</td>
<td>$K$</td>
<td>$V^+$</td>
</tr>
<tr>
<td>pwen</td>
<td></td>
<td>180</td>
<td>180</td>
<td>keV</td>
<td>$SWING$</td>
<td>$V$</td>
</tr>
</tbody>
</table>

$<1$ as low as possible, $>1$ as high as possible
5.4 Sensitivity of the responses

5.4.1 Total sensitivity in two points

The total sensitivity of the important parameters $V_{T-Nominal}$, $V_{T-Long}$ and $V_{PUNCH}$ is determined for two points: One point determined by the target criterium and the other point determined by the combination of the target and sensitivity criterium. All input factors are varied according to the variance in Tab. 2.3.

In Fig. 5.3, Fig. 5.4 and Fig. 5.5 the $p$-channel output parameter sensitivity is given in two points. For a $p$-channel transistor $V_{PUNCH}$ should be negative to realize the specification for $I_{OFF}$, and positive for the $n$-channel.

The mean long channel $V_T$ in Fig. 5.3 shifts if the criterium changes; The change in the standard deviation is minimal. The point based on the target criterium is therefore better for $V_T$, but can be worse for other responses.

For the nominal $V_T$ in Fig. 5.4 the combined criterium better achieves the specification. The nominal $V_T$ is larger than 0.5V for all variations around the mean, if the combined criterium is used.

For $V_{PUNCH}$ both points satisfy the constraint. Both points result in a negative voltage, meaning that $I_{OFF}$ satisfies the constraint. The difference in the deviation is more clear for this response than the other two. The combined criterium results in a lower deviation of $V_{PUNCH}$. 
**Fig. 5.3** Long $V_T p$-channel in 2points

**Fig. 5.4** Nom. $V_T p$-channel in 2points
Results of the optimized process

Fig. 5.5 $V_{PUNCH}$ p-channel in 2points
In Fig. 5.6, Fig. 5.7 and Fig. 5.8 the n-channel output parameter sensitivity is given in two points.

The mean long channel $V_T$ in Fig. 5.6 shifts if the criterium changes; There is a big change in the standard deviation. The point based on the target criterium has a large deviation.

For the nominal $V_T$ in Fig. 5.7 both points satisfy the specification: The nominal $V_T$ is larger than 0.37V for all variations around the mean. But the point determined by the combined criterium gives a much lower deviation around the mean and is preferred above the target criterium.

For $V_{PUNCH}$ both points are bad. Both points result in a negative voltage, meaning that $I_{OFF}$ is higher than the specified value. The difference in the deviation is not significant.

![Fig. 5.6 Long $V_T$ n-channel in 2points](image)
5.4.2 Sensitivity per input factor

In the paragraph above the sensitivity is determined by varying all input factors. In this paragraph the influences per input factor are viewed. This is done only for the most important output factor $V_T$. Via these variations the most important input factors can be shown. Fig. 5.9 and Fig. 5.10 show the $p$ channel variation per input factor for the optimum point based on the target criterium and the combined criterium. Fig. 5.11 and Fig. 5.12 show how this is done for the $n$ channel.
Fig. 5.9 p-ch Sens. per input/target optimum
Results of the optimized process

Fig. 5.10  $p$-ch Sens. per input/combined optimum
### Simulation Target Optimum \( \sigma_{\text{cell}=30 \text{ cm}^4} \)

<table>
<thead>
<tr>
<th>Mean Value</th>
<th>Standard Dev.</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.090E-01</td>
<td>6.189E-04</td>
<td>-7.830E-02</td>
<td>1.316E+01</td>
</tr>
</tbody>
</table>

### Simulation Target Optimum \( \sigma_{\text{cell}=4 \text{ keV}} \)

<table>
<thead>
<tr>
<th>Mean Value</th>
<th>Standard Dev.</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.090E-01</td>
<td>7.795E-03</td>
<td>2.185E-01</td>
<td>-9.933E-02</td>
</tr>
</tbody>
</table>

### Simulation Target Optimum \( \sigma_{\text{cell}=0.6 \text{ keV}} \)

<table>
<thead>
<tr>
<th>Mean Value</th>
<th>Standard Dev.</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.090E-01</td>
<td>3.630E-03</td>
<td>2.835E-01</td>
<td>1.474E-01</td>
</tr>
</tbody>
</table>

### Simulation Target Optimum \( \sigma_{\text{cell}=10 \text{ nm}} \)

<table>
<thead>
<tr>
<th>Mean Value</th>
<th>Standard Dev.</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.093E-01</td>
<td>2.676E-03</td>
<td>-1.349E+00</td>
<td>2.216E+00</td>
</tr>
</tbody>
</table>

Fig. 5.11  *n-ch Sens. per input/target optimum*
Results of the optimized process

Fig. 5.12  n-ch Sens. per input/combined optimum
5.5 The optimum for the real transistor

The accuracy of the input setting in the optimum depends on the accuracy of the model as determined with the simulators, and the accuracy which the simulators accomplish in describing the real process. It is important to know how this optimum can be used in the real process.

5.5.1 Accuracy of the model

The accuracy of the model is determined by the parameter $R^2$ declared in Tab. 4.3 and Tab. 4.4. The accuracy depends on the amount of simulations and the order of the model used. The accuracy is good as can be seen in the table.

5.5.2 Accuracy of the real process

In order to assure an accurate model related to the real CMOS process, the MINIMOS parameters of the 0.35μm CMOS process were used [7]. The 0.35μm process parameters were extracted by fitting a model to the real fabricated 0.35μm CMOS process. To determine the real accuracy of the simulated 0.25μm CMOS process the simulations have to be compared with real experiments.

5.5.3 The simulated optimum used in the real process

The simulated optimum, as described in Tab. 4.3 and Tab. 4.4, can directly be used as a starting value for the real process.

The plots in Fig. 5.1 and Fig. 5.2 can be used to interpret the relation between the input and output parameter to determine extra experimental points to check the relation. These plots give a trend of the expected relation.

The sensitivity plots in Fig. 5.3 to Fig. 5.12 can be used as an indication to which output factors are very sensitive for process steps.
5.6 Conclusion

In theory an optimum with all responses satisfying the constraints is impossible with the above model. The required change in $V_T$ is not acceptable.

The optimized process is determined within the accuracy of the simulations. The optimum can not be found to satisfy all output specifications, but is situated within a larger range of specifications.

The accuracy of the simulated optimum is restricted to the accuracy of the individual simulation programs and the model determined between the input and the output parameters.

To get the best setting for the MINIMOS device old MINIMOS parameters are used which were best fitted to the real 0.35μm CMOS process.

If the model satisfies to the real device the optimum is not possible in practice.

From the model the following can be extracted:
1. Some input parameters are very critical. ($V_T$ implantation, the thickness of the oxide, the APT implantation and $L_{eff}$). This follows from the sensitivity analysis and the model coefficients.
2. There are contradictory interests of some input parameters: $V_T$ and $I_{SAT}$ for the $n$-channel transistor and $V_T$ and $I_{OFF}$ for the $p$-channel transistor.
3. The specification is very stringent.

The model can be used for two reasons:
1. To determine the optimal parameter setting before doing real experiments.
2. To determine the relation between the in and output parameters functional.
6 Conclusion

The new software NORMAN and DEBORA has been used to optimize the 0.25μm CMOS process. Statistical methods like the Response Surface Method are used to generate experimental designs like factorial design and Central Cubic Faced design. Another special design technique that is used in NORMAN is the Target Oriented Design. These designs are used to reduce the amount of simulations needed to find a model describing the input process parameters and the device parameters.

NORMAN and DEBORA automates the CMOS simulation process. It works fast and it reduces the time needed before a model is extracted. This is caused by reducing the amount of simulations needed by having the ability to select a design method. DEBORA uses the model to automatically optimize the process to performance and reproducibility; These two criteria, that can be used separately and combined, are important for process optimization.

The ability to do transformations in NORMAN is incorporated, but the transformed variables cannot yet be used in DEBORA, which should be incorporated. Another negative point of the software is that although it works automatically, much handling on files is needed, which makes it user unfriendly. A great deal of expertise of individual simulation programs is needed to work with NORMAN and DEBORA.

The statistical method RSM is a good method to generate a model describing the input and output parameters. It is also a method to investigate the resulting model, to get accurate results. The design methods drastically reduce the amount of simulations needed to describe the model.

There are many possible designs, but a great expertise is needed to use them. They would better be used by just selecting the right design by the right kind of model.

The result of the optimization of the 0.25μm CMOS process is a setting of the input parameters where the responses are in a defined output range. This range is larger than the given specifications of the responses. The output range is enlarged because otherwise no point in the input range can be found. The resulting optimum is found by using the target criterium. In this optimum a sensitivity analysis is done, that shows the process sensitivity and the order of the input parameters. The input setting can be used as starting parameters for the real process. These real experiments can be used to check the accuracy of the model and to improve it.
1. Gaston, G.J. and A.J. Walton
   THE INTEGRATION OF SIMULATION AND RESPONSE SURFACE METHODOLOGY FOR THE OPTIMIZATION OF IC PROCESSES.

2. Walton, A.J. and M. Fallon, M.I. Newsam
   USING RSM TECHNIQUES TO CONTOUR PLOT RESPONSE DISTRIBUTIONS OF SEMICONDUCTOR PROCESSES.
   In: Proc. 1994 ESSDERC, P. 829-832

   DEVICE PARAMETER CHANGES CAUSED BY MANUFACTURING FLUCTUATIONS OF DEEP SUBMICRON MOSFET'S.

   ADVANTAGE OF THE METHODOLOGY USING DOE AND SIMULATION FOR OPTIMISING ADVANCED TECHNOLOGIES APPLICATION TO A REAL 0.35µM PMOS ARCHITECTURE.
   In: Proc. 1994 ESSDERC, P. 841-844

   THE IMPACT OF SCALING ON HOT-CARRIER DEGRADATION AND SUPPLY VOLTAGE OF DEEP-SUBMICRON NMOS TRANSISTORS.
   In: Proc. 1991 IEDM, P 537-540

   MOS MODEL 9.
   Eindhoven: Unclassified report NL-UR 003/94 Philips Electronics N.V.

   A 0.35µM CMOS TECHNOLOGY OPTIMIZED FOR LOW-POWER APPLICATIONS
   In: Proc. 1994 ESSDERC, P. 189-190

   INFLUENCE OF HIGH SUBSTRATE DOPING LEVELS ON THE TRESHOLD VOLTAGE AND THE MOBILITY OF DEEP-SUBMICRON MOSFET'S
   IEEE TRANSACTIONS ON ELECTRON DEVICES, Vol. 39 (1992), NO. 4, P. 932-938

9. Box, G.E.P. and N.R. Draper
   EMPIRICAL MODEL-BUILDING AND RESPONSE SURFACE.

    PROCESS TECHNOLOGY OPTIMIZATION USING AN INTEGRATED PROCESS AND DEVICE SIMULATION SEQUENCING SYSTEM
    MICROELECTRONIC ENGINEERING, Vol. 19 (1992), P. 507-510

11. Cartuyvels, R
    NORMAN THESIS
    Leuven, 1993, P. 34-38

12. Booth, R and R Cartuyvels, L. Dupas
    NORMAN & DEBORA, Leuven, 1994
This appendix visualizes the one dimensional doping profile under the gate, simulated by the process simulator SUPREM3. The dope concentration is plotted as a function of the depth of the transistor. In this case the p-channel transistor has 3 Phosphorus implantations and 1 Arsenic implantation:
1. Well (P)
2. Channel Stopper (P)
3. Anti Punch Trough (P)
4. Threshold Voltage adjust (As)
The last plot is the concentration after a temperature step. The first 0.2 microns is poly Silicium. The gate oxide is not visible.
<table>
<thead>
<tr>
<th>Number of runs N</th>
<th>Number of variables k</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2^{12}</td>
</tr>
<tr>
<td>8</td>
<td>2^{24}</td>
</tr>
<tr>
<td>16</td>
<td>2^{48}</td>
</tr>
<tr>
<td>32</td>
<td>2^{96}</td>
</tr>
<tr>
<td>64</td>
<td>2^{192}</td>
</tr>
<tr>
<td>128</td>
<td>2^{384}</td>
</tr>
</tbody>
</table>

**Appendix 2: Two-level fractional factorial design for k variables and N runs**
Appendix 3: NORMAN input file

C Kwartmicron simulatie met SD_nplus.tsup-files CMOS-example
C ben

@FACTORS ! nominal low high
sacox 0.010 .005 .016 ! sacrificial oxide mu

pwdo 1E13 1.0e12 1.4e13! p well dose N channel cm-2
pwen 180 170 190 ! p well energy N channel keV
#csenaptncdo 1.0E12 6e11 6e12 ! Channel Stop N channel dose cm-2
#csenaptncen 90 60 90 ! Channel Stop N channel energy keV
!aptncdo 2.5E12 0.5e12 5.0e12! Anti Punch Trough N channel dose cm-2
!!!aptncen 66 40 70 ! 80MAXbij tr.boron Anti Punch Though

N channel energy cm-2
# vtncdo 6.5E12 6e12 9e12 ! Vt Implantation N channel dose cm-2
# vtncen 16 5 50 ! Vt Implantation N channel energy keV
Twell 850 ! well drive degrees-C
# dox 0.006 0.0058 0.0062! gate oxide thickness mu
Tox 850 ! Temperature oxide growth degrees-C
# varl 0.0 -0.05E -4 0.05E-4

@SEQUence

!EXP=FRA CONF=2345;1346;1237;1248
!EXP=pbd
!EXP=FILE NAME=mydesign BEGIN=80 RESume=7 LAST=15
!EXP=FRA CONF=1235;2346 BEGIN=18
!exp=nom BEGIN=80
!EXP=FULL
!EXP=DIAgonal NPTS=5 BEGIN=80

EXP=ccf RESUME=15 BEGIN=1

!=================================================
! generate suprem3 input files
! ==================================================

PROcess
sys echo process done

let Llang=10E-4 + $varl$
let Lkort=0.355E-4 + $varl$
let Lioff=0.305E-4 + $varl$
echo $Llang$ $Lkort$ $Lioff$

! ===========
! suprem3 run
! ===========
Appendix 3: NORMAN input file

! make supre3-inputfiles
SYS cp X1_$EXP$.INP process1$EXP$.sup3inp
! run supre3
sys rsup3 -c -t 10 process1$EXP$
sys echo supremrun done

! make supmin-inputfile
SYS cp process1$EXP$.sup330 process1$EXP$.sup3

! minimos doping profile

!=======================
! start supmin
SYS supmin 2 SD_nplus.tsup process1$EXP$.sup3 process1$EXP$.dope
SYS echo smrunnes done

! minimos4 run

LET TOX=$dox*$1.0-4
!mmos4 Deck-input-files
make deck=device1.INP
make deck=device3.INP
make deck=device5.INP
make deck=device7.INP
make deck=device8.INP
make deck=device11.INP
make deck=device13.INP

!mmos4 process-Deck-input-files
SYS cp device1.INP device1$EXP$.INP
SYS cp device3.INP device3$EXP$.INP
SYS cp device5.INP device5$EXP$.INP
SYS cp device7.INP device7$EXP$.INP
SYS cp device8.INP device8$EXP$.INP
SYS cp device11.INP device11$EXP$.INP
SYS cp device13.INP device13$EXP$.INP

!making of the job-files
SYS echo INPUT >rjob1_file
SYS echo $EXP$ >EXP_num
SYS echo minmos device1 process1 >target
SYS cat EXP_num target1 awk -f bewerk_een.awk >>rjob1_file
SYS cat deelminimos >>rjob1_file
SYS cat EXP_num target1 awk -f bewerk_twee.awk >>rjob1_file

SYS echo INPUT >rjob3_file
SYS echo minmos device3 process1 >target
Appendix 3: NORMAN input file

```
SYS cat EXP_num target | awk -f bewerk_een.awk »rjob3_file
SYS cat deelminimos        »rjob3_file
SYS cat EXP_num target | awk -f bewerk_twee.awk »rjob3_file

SYS echo INPUT    »rjob5_file
SYS echo minmos device5 process1 »target
SYS cat EXP_num target | awk -f bewerk_een.awk »rjob5_file
SYS cat deelminimos        »rjob5_file
SYS cat EXP_num target | awk -f bewerk_twee.awk »rjob5_file

SYS echo INPUT    »rjob7_file
SYS echo minmos device7 process1 »target
SYS cat EXP_num target | awk -f bewerk_een.awk »rjob7_file
SYS cat deelminimos        »rjob7_file
SYS cat EXP_num target | awk -f bewerk_twee.awk »rjob7_file

SYS echo INPUT    »rjob8_file
SYS echo minmos device8 process1 »target
SYS cat EXP_num target | awk -f bewerk_een.awk »rjob8_file
SYS cat deelminimos        »rjob8_file
SYS cat EXP_num target | awk -f bewerk_twee.awk »rjob8_file

SYS echo INPUT    »rjob11_file
SYS echo minmos device11 process1  »target
SYS cat EXP_num target | awk -f bewerk_een.awk »rjob11_file
SYS cat deelminimos        »rjob11_file
SYS cat EXP_num target | awk -f bewerk_twee.awk »rjob11_file

SYS echo INPUT    »rjob13_file
SYS echo minmos device13 process1  »target
SYS cat EXP_num target | awk -f bewerk_een.awk »rjob13_file
SYS cat deelminimos        »rjob13_file
SYS cat EXP_num target | awk -f bewerk_twee.awk »rjob13_file

SYS echo start "njobfile" which starts rjob1_file..rjob6_file
!jobfile starts rjob1_file..rjob6_file in one shell in the background
!and waits until all background jobs in the particular shell
!are stopped

SYS njobfile
SYS echo all jobs finished
!

=================================
! parameter extraction
=================================

SYS rm -f parexrun
SYS echo parex device1$EXP$ 9 3 0 \\
>parexrun
```
SYS echo parex device3$EXP$.parex Ddevice1$_$EXP$.EXT
SYS cp device3$EXP$.parex Ddevice3$_$EXP$.EXT
SYS cat device5$EXP$.data | grep -e '2.500D+00' | awk -f isat.awk > mrt
SYS cat mrt >> Ddevice3$_$EXP$.EXT !isat
SYS cat device7$EXP$.data device8$EXP$.data | awk -f swing.awk > mrt
SYS cat mrt >> Ddevice3$_$EXP$.EXT !subthresholdswing
SYS cat device1$_$EXP$.data | grep -e '2.750D+00' | awk -f ioff.awk > mrt
SYS cat mrt >> Ddevice3$_$EXP$.EXT !ioff
SYS cat device13$EXP$.data | awk -f vpunch.awk > rnrt
SYS cat mrt » Ddevice3$_$EXP$.EXT !vpunch
SYS rm -f *.INP
SYS rm -f *.sup3inp *.minkop *.sup3cons *.sup3 *.sup330 *.log
SYS rm -f *.OUT *.sup3out rjob*_file
SYS rm -f *.sup3s
SYS rm -f device5* device8* device9* device11* device13* device7*
SYS rm -f job-* *dia
SYS rm *.data
SYS rm *.parex *.job
SYS rm -f *.dump *.mos*.list*
SYS rm -f *.ref ftn*
SYS rm -f *.bin
SYS rm -f *.out
SYS rm -f *.dope
end
!end EXPerimental loop

!COLLECT !make files of the extracted parameters
SYS $D$collect DESIGN Ddevice1_001.EXT Pdevice1_VTO VTO
SYS $D$collect DESIGN Ddevice1_001.EXT Pdevice1_BETA BETA
SYS $D$collect DESIGN Ddevice1_001.EXT Pdevice1_TETAA TETAA
SYS $D$collect DESIGN Ddevice1_001.EXT Pdevice1_K K
SYS $D$collect DESIGN Ddevice3_001.EXT Pdevice3_VTO VTO
SYS $D$collect DESIGN Ddevice3_001.EXT Pdevice3_BETA BETA
SYS $D$collect DESIGN Ddevice3_001.EXT Pdevice3_TETAA TETAA
SYS $D$collect DESIGN Ddevice3_001.EXT Pdevice3_K K
SYS $D$collect DESIGN Ddevice3_001.EXT Pdevice3_ISAT ISAT
SYS $D$collect DESIGN Ddevice3_001.EXT Pdevice3_SWING SWING
SYS $D$collect DESIGN Ddevice3_001.EXT Pdevice3_IOFF IOFF
Appendix 3: NORMAN input file

SYS $D$collect DESIGN Ddevice3_001.EXT Pdevice3_VPUNCH VPUNCH

! ! The suprem3 processing steps
! +++++++++++++++++++++++++++++++++++++++++++++++++++
! cross section #1: LONG NCHANNEL
! cross section #2: LONG PCHANNEL
! +++++++++++++++++++++++++++++++++++++++++++++++++++
!
This section is company restricted

!
! Minimos Input decks
!
!
! long-channel NMOSFET
!
@DECK NAME=device1.INP
*NMOST - norman example
DEVICE CHANNEL=N GATE=NPOLY TINS=$TOX$ W=10E-4 + L=$L$lang$ DGAP=0.5E-4 SGAP=0.5E-4
BIAS UG=2.5 UD=0.1 UB=0.0
*STEP NG=10 DG=-0.25
STEP NG=5 DG=-0.25 NB=2 DB=-1.5
PROFILE FILE=2-D
MOBILITY MC=0.85 MT=1.1 MR=1.0
OPTION MODEL=2-D *QMCOR=YES GRIDFREEZE=YES *********UL=300E-8
OUTPUT PSI=YES*NONE=YES
END BIN=YES ERR=1.E-4
!
! short-channel NMOSFET
! even punch through!!
@DECK NAME=device3.INP
*NMOST - norman example
DEVICE CHANNEL=N GATE=NPOLY TINS=$TOX$ W=10E-4 + L=$L$kort$ DGAP=0.5E-4 SGAP=0.5E-4
BIAS UG=2.5 UD=0.1 UB=0.0
*STEP NG=10 DG=-0.25
STEP NG=5 DG=-0.5 NB=2 DB=-1.5
PROFILE FILE=2-D
MOBILITY MC=0.85 MT=1.1 MR=1.0
OPTION MODEL=2-D *QMCOR=YES GRIDFREEZE=YES *********UL=300E-8
OUTPUT PSI=YES*NONE=YES
END BIN=YES ERR=1.E-4
!
! short-channel NMOSFET SATURATION CURRENT ISAT
Appendix 3: NORMAN input file

70

! @DECK NAME=device5.INP
*NMOST - norman example
DEVICE CHANNEL=N GATE=NPOLY TINS=$TOX$ W=10E-4
+ L=$Lkort$ DGAP=0.5E-4 SGAP=0.5E-4
BIAS UG=2.5 UD=2.5 UB=0.0
PROFILE FILE=2-D
MOBILITY MC=0.85 MT=1.1 MR=1.0
OPTION MODEL=2-D PHYS=NO *QMCOR=YES *************UL=300E-8
OUTPUT NONE=YES
END BIN=NO ERR=1.E-4

! short-channel NMOSFET subthresholdswing
!
@DECK NAME=device7.INP
*NMOST - norman example
DEVICE CHANNEL=N GATE=NPOLY TINS=$TOX$ W=10E-4
+ L=$Lkort$ DGAP=0.5E-4 SGAP=0.5E-4
BIAS UG=0.5 UD=0.1 UB=0.0
PROFILE FILE=2-D
MOBILITY MC=0.85 MT=1.1 MR=1.0
OPTION MODEL=THRES PHYS=NO CUR=2.82E-7 *QMCOR=YES ****UL=300E-8
OUTPUT NONE=YES
END BIN=NO ERR=1.E-5

! short-channel NMOSFET subthresholdswing
!
@DECK NAME=device8.INP
*NMOST - norman example
DEVICE CHANNEL=N GATE=NPOLY TINS=$TOX$ W=10E-4
+ L=$Lkort$ DGAP=0.5E-4 SGAP=0.5E-4
BIAS UG=0.5 UD=0.1 UB=0.0
PROFILE FILE=2-D
MOBILITY MC=0.85 MT=1.1 MR=1.0
OPTION MODEL=THRES PHYS=NO CUR=2.82E-8 *QMCOR=YES ****UL=300E-8
OUTPUT NONE=YES
END BIN=NO ERR=1.E-5

! short-channel NMOSFET SATURATION CURRENT IOFF
!
@DECK NAME=device11.INP
*NMOST - norman example
DEVICE CHANNEL=N GATE=NPOLY TINS=$TOX$ W=10E-4
+ L=$Lioff$ DGAP=0.5E-4 SGAP=0.5E-4
ABIASS    UG=0.0  UD=2.75  UB=0.0
PROFILE   FILE=2-D
MOBILITY   MC=0.85  MT=1.1  MR=1.0
OPTION     MODEL=2-D  PHYS=NO  *QMCOR=YES  ***************UL=300E-8
OUTPUT     NONE=YES
END        BIN=NO   ERR=1.E-4

short-channel NMOSFET Vpunch through
@DECK NAME=device13.INP
*NMOST - norman example
DEVICE     CHANNEL=N  GATE=NPOLY  TINS=$TOX$  W=10E-4
+ L=$Lioff$  DGAP=0.5E-4  SGAP=0.5E-4
BIAS       UG=0.5    UD=2.75  UB=0.0
PROFILE    FILE=2-D
MOBILITY   MC=0.85  MT=1.1  MR=1.0
OPTION     MODEL=THRES  PHYS=NO  CUR=1E-10  *QMCOR=YES  ****UL=300E-8
OUTPUT     NONE=YES
END        BIN=NO   ERR=1.E-4
@END      !end of Norman input file