MASTER

Improving low-voltage integrated power devices
a simulation and experimental based study

Toonen, J.

Award date:
2019

Faculty Applied Physics
Nanoscience & technology
Photonics and Semiconductor Nanophysics

February 2018 - February 2019

Improving low-voltage integrated power devices

a simulation and experimental based study

Jelke Toonen

Supervisors:
Ir. A. Mels
Dr. Ir. H.G.A. Huizing
Prof. Dr. Koenraad

In partial fulfilment of a master thesis

Improving low-voltage integrated power devices: a simulation and experimental based study

Master of science thesis

Date: 18 January 2019

Author: J. Toonen
MSc student Applied Physics
Eindhoven University of technology
Faculty of Applied Physics
Department of Nanoscience & Technology
Research group Photonics and Semiconductor Nanophysics

Company supervisors: Ir. A. Mels
Technical Director High Voltage and Power
NXP Semiconductors Nijmegen
Front End Innovation
High Voltage & Power Technology Nijmegen

Dr. Ir. H.G.A. Huizing
Manager High Voltage and Power
NXP Semiconductors Nijmegen
Front End Innovation
High Voltage & Power Technology Nijmegen

University supervisor: Prof. Dr. P. Koenraad
Eindhoven University of technology
Faculty of Applied Physics
Department of Nanoscience & Technology
Research group Photonics and Semiconductor Nanophysics

Version: Public

Preface

This thesis on “Improving low-voltage integrated power devices: a simulation and experimental based study” marks the end of my graduation project, which completes my Master of Science in Applied Physics offered at the Eindhoven University of Technology (TU/e). The research presented here is performed in cooperation with NXP Semiconductors located in Nijmegen. This work, as well as my personal development, would have not been possible without the support I received from both NXP and the TU/e.

First of all, I would like to thank Dr. Ir. H.G.A. Huizing for offering me the opportunity to conduct this innovative investigation within the High Voltage and Power Technology group at NXP in Nijmegen. Surrounded by top-of-the-line engineers, scientists and managers, working in this group really pushed my boundaries and served as an ideal learning environment at the verge of my own career. In addition, it also made me develop a passion for the field of device physics, which in my opinion is one of the greatest examples where fundamental understanding of the nature can lead to products used all over the world.

Secondly, I would like to thank Ir. A. Mels for his help whenever it was necessary. As my daily advisor, he had a crucial role in the development of this work. His fast thinking and extensive knowledge of the semiconductor industry always pointed me in the right direction.

Furthermore, I would like to thank Ir. I. Emmerik - Weijland for all her support related to measurements and the interpretation of them. Moreover, the weekly discussions we had were always very fruitful for my deeper understanding of device physics.

Also, I would like to thank all my colleagues in general within the High Voltage and Power Technology group for providing me an enjoyable working environment. Moreover, the wide variety of expertise’s available in this group was also a very valuable addition. In particular, Dr. S. Sque and Dr. S. Mehrotra greatly supported me with my simulations, Ir. P. Boos helped me to better understand the processing aspect of semiconductor devices and Ir. E. van der Sar kept a clear overview of the project as well as my role in it.

Finally, I would like to thank my mentor from the TU/e, Prof. Dr. P. Koenraad, for inquiring his contacts for arranging this internship at NXP and for the subsequent supervision of it.

Eindhoven, January 2019

Summary

Modern society is more and more dependent on electrical appliances for all kinds of purposes such as comfort, healthcare and communication, putting increasing demands on the performance of Integrated Circuits (IC's). Besides the standard Metal-Oxide-Semiconductor Field-Effect Transistors (MOSFET's) for (low voltage) CMOS (Complementary Metal-Oxide-Semiconductor) technology, IC's also contain integrated power devices to handle the higher voltages. Such devices are mostly asymmetrical in nature: they have a MOSFET channel region required for switching and a lowly doped extension of the drain (referred to as drain-extension) to handle the majority of the voltage drop. Integrated power devices are typically used as a switch, requiring that they can handle large currents in the on-state with minimal static power dissipation and that they can sustain a large voltage in the off-state with minimal leakage current. Two important FOM's (figure-of-merits) that capture these requirements are the specific on-state resistance ($R_{ON}A$) and the off-state breakdown voltage (BV), between which a significant trade-off exists (in 1D: $R_{ON}A \propto BV^{11/3}$ and in 2D: $R_{ON}A \propto BV^{7/3}$) [1].

The objective of this thesis is to improve integrated power devices in the existing NXP technology platform to obtain superior $R_{ON}A$-BV trade-off with respect to state-of-the-art devices in the range up to 50V off-state breakdown. For this purpose, two novel concepts, namely oxygen-inserted layer (OIL)- and contact field plate (CFP) technology, are explored by both simulation and experiment. In the OIL approach, a thin film (~10 nm) of reengineered silicon that incorporates partial monolayers of oxygen is buried close to the silicon surface. For CMOS applications, it is demonstrated that this decreases the resistance by an increase in channel mobility [2]–[7]. In the CFP approach, a contact module which is normally used for connecting the source, drain and body, is used as a field plate on top of the drain-extension [8]. This can be used to both optimize the $R_{ON}A$ and BV, while no additional masks or process changes (i.e. costs) are needed [9], [10]. For both concepts a full wafer lot was set up, in which the OIL concept is applied to the full range of devices available in the technology platform whereas the CFP concept was only suitable for the 12V device. In addition, for both concepts also simulations are set up in Synopsys Sentaurus™ TCAD (Technology Computer-Aided Design), which for the OIL concept are mainly focused on properly integrating the film and for the CFP concept on obtaining electrical characteristics.

On the wafer lot with the OIL experiments, extensive electrical analysis's have been performed. However, the results that are obtained (which are omitted in this public report for confidentiality reasons) are not conclusive yet. From the wafer lot with the CFP experiments, it became evident from electrical measurements and simulations that the devices were far from optimized. For the BV, the crucial parameter turned out to be the oxide thickness below the CFP ($t_{ox,FP}$), which in this wafer lot was believed to be significantly below the optimum. For the $R_{ON}A$, significant room for improvement is in reducing the lateral dimensions, especially that of the channel. Finally, to show how the CFP devices can be optimized, without using any additional masks (i.e. costs), optimizations are proposed by simulations for the initial 12V application as well as for higher voltages. For the initial (12V) application, the $R_{ON}A$ is reduced by ~40% resulting in a device that has the potential to beat state-of-the-art devices by ~20%. Using the same architecture, BV's up to ~35V can effectively be achieved by appropriate scaling of the drain-extension and $t_{ox,FP}$. Scaling effectively to higher voltages requires a newly proposed CFP architecture, in which multiple (rotated) contacts are placed in an array along the width of the device.

So, simulation results showed that utilizing the CFP concept and subsequent optimization provides a way to beat current state-of-the-art integrated power devices (for BV's<50V) in the existing NXP technology platform. Moreover, if the OIL works as previously published [2]–[7], this concept has the potential to add additional improvement. In order to demonstrate this on silicon, follow-up cycles are needed for both concepts.

Abbreviations

CMOS	Complementary Metal-Oxide-Semiconductor
CMU	Capacitance-Measure-Unit
DIBL	Drain Induced Barrier Lowering
DIELER	Dielectric
DG	Density Gradient
DOS	Density Of States
DUT	Device Under Investigation
FET	Field-Effect Transistor
FP	Field Plate
FOM	Figure-Of-Merit
FA	Free-Analog
GIDL	Gate Induced Drain Leakage
HCI	Hot Carrier Injection
HH	Heavy Hole
IC	Integrated Circuit
II	Impact Ionization
JFET	Junction Field-Effect Transistor
LDD	Lightly Doped Drain
LDMOS	Laterally Diffused Metal-Oxide-Semiconductor
LH	Light Hole
MOS	Metal-Oxide-Semiconductor
MOSFET	Metal-Oxide-Semiconductor Field-Effect Transistor
MST	Mears Silicon Technology
NPT	Non-Punch Through
OIL	Oxygen-Inserted Layer
RESURF	Reduced SURface Field
PM	Partial Monolayer
PCM	Process Control Monitoring
PT	Punch Through
SIMS	Secondary Ion Mass Spectroscopy
SiProt	Silicon Protection layer of oxide
SMU	Source-Measure-Unit
SRH	Schockley-Read-Hall
SO	Spin-Orbit
SOA	Safe Operating Area
SSRW	Super-Steep-Retrograde-Well
STI	Shallow Trench Isolation
TCAD	Technology Computer-Aided Design
TED	Transient Enhanced Diffusion
TEM	Transmission Electron Microscopy

Contents

Chapter 1: Introduction and motivation

The High Voltage and Power Technology group within NXP is responsible for developing and enabling new integrated high voltage and power semiconductor technologies. Besides the design of devices, this group is also responsible for the industrialization. Many device types over a wide application range are already delivered by this group which compete at the highest level of power technology in the market. In order to stay one of the market leaders in power technology, the devices should be continuously improved and/or new novel device concepts should be designed. Therefore, a long-term innovation project is initiated within this group. The goal of this project is to propose/demonstrate device concepts that outperform current state-of-the-art integrated power devices in the range up to 50V off-state breakdown, which also have the potential to be integrated in the existing NXP technology platform. Here the specific on-state resistance (on-state resistance times area on the wafer) and the off-state breakdown voltage are the key design targets. In order to obtain a clear overview of possibilities to improve these key design targets, a literature study is conducted. From this literature study, some promising concepts are selected and further explored by simulation and/or experiment in this work.

In this chapter, first the context for the need of this research will be provided. Thereafter, it will be discussed which concepts are currently known from literature to improve integrated power devices and which are particularly explored in this work. Lastly, the objective will be defined and the outline of this thesis will be shown.

1.1 Market demand

The continuous integration of sophisticated electronic technologies like mobile phones and computers in today's modern society, puts more and more demands on current technology. Since the heart of most electronic devices consists of an Integrated Circuit (IC), these must be continuously improved. In order to increase the capacity and the functionality of these IC's, the number of transistors per IC must be increased. One way to do this, is to shrink the size of the transistor. Not only this decreases the cost per transistor, but also lowers the power consumption while the speed and memory capacity go up [11]. Therefore, miniaturization is nowadays without a doubt the largest research interest within the semiconductor industry. Integration of inherently novel 2D/3D device architectures, materials with superior characteristics and/or band structure engineering have gradually become a requirement to keep up with the surprisingly long-lasting prediction set by Moore in 1965 [12]. Yet, for the power semiconductor industry, some of these innovative design trends already became necessary long before their low power counterparts. The range of power semiconductor devices is generally very broad and contains everything from the high power (>10 MW) low switching speed thyristors, the mid-range (1 kW - 1 MW) insulated gate bipolar transistors, to the low power (<1 kW) high switching speed double diffused MOSFET (Metal-Oxide-Semiconductor Field-Effect Transistor) [9], [13]. Besides the applicable range, power devices can further be subdivided by their orientation. Discrete power devices are vertically oriented such that their drain contact is at the bottom. Integrated power devices on the other hand, are laterally oriented demanding their drain contact at the top. The focus in this research is on integrated power devices in the lower end of the power spectrum, covering the standard (symmetric) MOSFET's suited for low voltage applications and the (asymmetric) Laterally Diffused Metal-Oxide-Semiconductor (LDMOS) devices suited for higher voltage applications.

Power devices are generally used as switches in power conversion systems, where by periodically switching the input current from the source the waveform can be converted and matched to the requirements of the load [10], [13]. An important requirement for such a switch, is that it can sustain a large voltage with minimal leakage current when it is open (off-state). This is generally done by using a so-called drain-extension, which is an extension between the channel and the drain of a MOSFET structure designed to support a large voltage drop. Therefore, an important figure-of-merit (FOM) used to classify the blocking capability of power devices, is the off-state breakdown voltage (BV). Another important requirement for a switch, is that it must be able to handle large currents when it is closed (on-state) with minimal static power dissipation [10], [13]. Therefore, the on-state resistance (R_{ON}) between the terminals (i.e. the source and drain), must be as low as possible. The specific on-state resistance ($R_{ON}A$), where A is the device area on the wafer, additionally takes the device dimensions into account [9]. As this indirectly includes the costs of on wafer implementation, $R_{ON}A$ is typically used as another important FOM for power devices [9], [14].

So clearly, in optimizing power devices, the BV and $R_{ON}A$ are the main two key design targets. These design targets however, are both related to the doping concentration and to the length of the drain-extension. This results in that $R_{ON}A$ increases super linearly with the BV, in particular, for 1D LDMOS devices it can be approximated that $R_{ON}A$ is proportional to $BV^{11/3}$ [1]. This illustrates that it is extremely hard to improve one of the two without compromising the other. Too overcome this difficulty, many intriguing ideas and concepts for novel power devices are proposed over the years, some of which are already successfully industrialized. In order to explore and identify such novel concepts for LDMOS devices, a literature study is conducted to provide a clear overview for possible device concepts known from literature. The following section summarizes this literature study and elaborates on which novel device concepts are further explored in this work.

1.2 Novel device concepts

Based on an extensive literature study, an overview is made of novel concepts that can be used to improve the $R_{ON}A$-BV trade-off for integrated (LDMOS) power devices (Figure 1). The tree diagram is differentiated in concepts that mainly try to independently improve the BV or the $R_{ON}A$, although it must be noted that the two always have a (small) correlation. In the rest of this section, each topic will shortly be discussed.

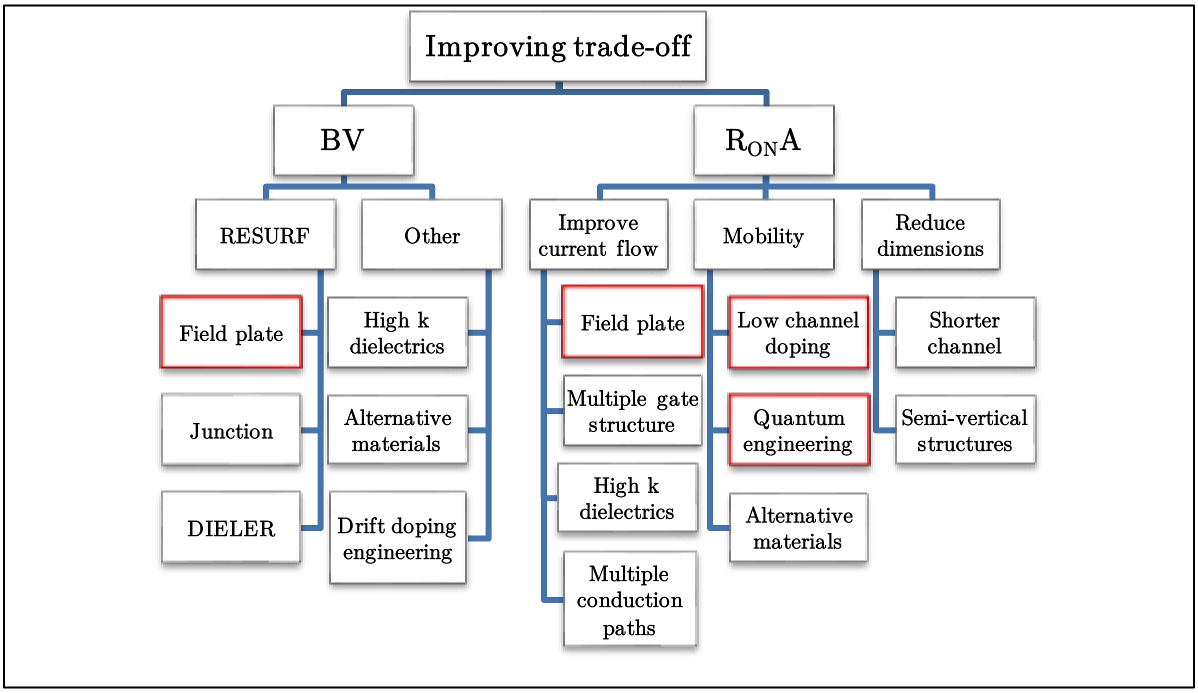

Figure 1: Schematic tree diagram of novel concepts that are present in the literature to improve the $R_{ON}A$-BV trade-off for (LDMOS) power devices. The tree is divided in concepts that mainly try to independently improve the BV or the $R_{ON}A$, although it must be noted that the two always have a (small) correlation. The red marked topics are the ones that are further explored in this work.

Probably the best-known approach to improve the BV, is by applying so-called RESURF (Reduced SURface Field) [15]. This approach is a collective name for concepts that optimize the lateral electric field distribution in the drain-extension using additional perpendicular fields. There are three typical variations of RESURF: field plate (FP)-, junction- and dielectric (DIELER) RESURF. Field plates are used on top and/or below the drain-extension, where they impose electrostatic boundary conditions. These come in many different variations, such as: extended gate [16], recessed gate [17], metal field plate [8], contact field plate [8], (semi-)resistive field plate [18] or poly field plate [8], [19], [20]. In junction RESURF, one or multiple semiconductor domains are implemented in the drain-extension with opposite doping type to obtain charge balance. Also here variations are possible, such as the number of alternating domains or the dimension in which the junctions are processed [21], [22]. The DIELER RESURF concept employs a structure where the semiconducting material in the drain-extension is interleaved by dielectric layers parallel to the current path. These dielectric layers induce an additional perpendicular capacitance which lowers the lateral electric field [23]–[25].

Other ways to improve the BV, aim to elevate the maximal allowed critical field or to reduce parasitic leakage mechanisms. High k dielectrics, instead of SiO_2, can be used to reduce gate (or FP) leakage due tunneling suppression [26]. Moreover, high k dielectrics can also be beneficial to improve FP- and DIELER-RESURF [27]. Alternative materials, such as SiC and GaN, have a higher critical field than silicon at which breakdown occurs [14], [28]. The maximum allowed critical field can also be elevated by smart engineering of the doping concentration in the drain-extension (ideally linearly increasing along the current direction) [29]. Furthermore, this can also be used to optimize the field distribution in RESURF applications [9], [10].

Reducing the $R_{ON}A$ by improving the current flow, aims to increase the number of available charge carriers and/or to increase the effective area through which the current flows. A FP can be used to accumulate charge in the drain-extension when a bias is applied, thereby increasing the number of available charge carriers [10]. Multiple gate structures (e.g. FinFET's) both increase the charge density and the effective area through which current flows [30], [31]. To further increase the accumulated charge from gates/FP's, high k dielectrics can be used for the insulating spacer layer to increase the capacitance [32]. Another way to increase the effective area through which current flows, is creating multiple conduction paths from the source to the drain. This can for example be achieved by implementing an additional gate in a trench at the source side, thereby creating an additional (vertical) channel [33].

Increasing the mobility for reducing the $R_{ON}A$, can generally be done by reducing scattering events and/or reducing the effective transport mass [34]. A way scattering can be reduced in the channel of MOSFET's, is by ensuring low doping in the channel. This can for example be accomplished by epitaxially growing undoped silicon on diffusion barrier layers, such as carbon doped silicon [35] or oxygen partial monolayers [2]–[7]. Reducing both the effective mass and scattering can be done by changing the quantum mechanical properties of the lattice. Generally, this can be achieved by the use of strain [36], [37] and/or quantization [3]–[7]. Obviously, obtaining better mobility can also be established by using alternative materials with superior transport properties with respect to silicon, such as Ge or GaAs [11], [38].

The last way to improve the $R_{ON}A$, is making the lateral dimensions of the device as small as possible. This is also the most effective way since it both decreases the resistance and the area. While the length of the drain-extension is needed to support a large voltage, the channel region is only used for switching and should be made as small as possible. Ways this can be done are for example using halo implants (i.e. extra doping near the source and/or drain [11], [39]) or using self-aligned source/drain contacts [40]. The area of the device can also be severely reduced by employing semi-vertical structures. This can for example be done by making the channel vertical by integrating the gate in a trench [41] or making part of the drain-extension vertical by integrating a trench in it [42]. This is also done for discrete (vertical) power devices [14], however is much harder for integrated power devices due to requirement of a top drain contact.

Although most of the above discussed concepts have been proven to improve integrated (LDMOS) power devices, industrialization generally remains challenging because of process difficulties and/or large costs. One of the concepts used in this work to improve a LDMOS device is the contact field plate (CFP), which is inspired by L. Wei et al. [8]. In this approach, a contact module which is normally used for connecting the source, drain and body, is placed on top of the drain-extension acting as a field plate. This was proven to improve the reliability, while no additional masks or process changes (i.e. costs) are needed [8]. Additionally, like all FP's, the CFP has also the potential to improve the $R_{ON}A$-BV trade-off which is the main interest in this work.

Another concept that is explored in this work, is oxygen-inserted layer (OIL) technology. This technique is also referred to as Mears Silicon Technology (MST®) [43] which is promoted by a company called Atomera. In this technology, a thin film (~10 nm) of reengineered silicon that incorporates partial monolayers of oxygen is buried close to the silicon surface. In doing this, it is demonstrated for CMOS (Complementary Metal-Oxide-Semiconductor) applications that the current performance and reliability of the device go up, while the power consumption and variability go down [2]–[7]. In particular, the main feature of interest is a decrease in resistance due to an increase in channel mobility. This enhancement can be contributed to both quantum engineering and low channel doping. For integrated power devices however, this technology is not demonstrated yet and will be examined for the first time in this work.

1.3 Objective and outline

The objective of this thesis is to improve integrated power devices in the existing NXP technology platform to obtain superior $R_{ON}A$-BV trade-off with respect to state-of-the-art devices in the range up to 50V off-state breakdown. For integrated power applications, the technology platform contains both normal MOSFET's suited for the lower voltages and LDMOS devices for the higher voltages. The two concepts that are being explored for improving the devices, OIL- and CFP technology, will be investigated both by simulation and experiment. In addition, the OIL concept is investigated for the full range of devices available in the technology platform (<50V) whereas the CFP concept only is applicable to the 12V Free-Analog (FA) LDMOS type of device.

This thesis is organized as follows:

Chapter 2 provides the essential theoretical background information that is needed to understand this thesis. This chapter will start with the key features of basic device physics, mostly related to the MOSFET. Thereafter, the special class of LDMOS devices as well as their traditional $R_{ON}A$-BV trade-off will be introduced. Lastly, the novel ways that are used to improve this trade-off will be explained in more detail.

Chapter 3 focuses on the methodology of the TCAD (Technology Computer-Aided Design) Sentaurus Synopsys^{TM} simulation tool used in this work. This simulation tool is used for simulating the processing of devices (e.g. implantation, diffusion and oxidation) as well as for simulating electrical characteristics. This chapter will cover: the most important process steps of NXP's technology platform, an example of a process simulation, the used physical models and the set-up of electrical simulations.

Chapter 4 discusses the experimental methodology that is used to perform electrical measurements on the devices. This includes an elaboration on the hardware setup (probe station and device analyzer) and its use, as well as on the measurement routines that are used to perform a variety of electrical measurements.

Chapter 5 reports on the OIL approach to improve integrated power devices from NXP's technology platform. However, due to confidentially reasons any detailed specifications/results are omitted in this public report. Instead, a summary is given of the proceedings concerning this approach to lower the $R_{ON}A$-BV trade-off. This includes an elaboration on how the wafer lot is set up, which devices are studied, which experiments are performed and how the results are analyzed.

Chapter 6 reports on the CFP approach to improve the 12V FA device from NXP's technology platform. First, it will be discussed what this device looks like with the CFP and what kind of variations are present on the wafer lot. Thereafter, results from electrical measurements will be shown and substantiated by simulations. Lastly, optimized versions will be proposed by simulation for the initial 12V application as well as for higher voltages.

Chapter 7 summarizes the most important conclusions reached in this work and reflects on whether the objective has been accomplished.

Chapter 8 provides an outlook for future work in which specific recommendations are done for potential follow-up work, including additional silicon cycles.

Chapter 2: Theoretical background

The semiconductor industry, which is responsible for the fabrication and design of semiconductor devices, became a profitable business around 1960 [1], [11], [14]. Since that time, it has evolved to a major global industry of almost half a trillion US dollars [44]. One of the reasons it could become so large, is the solid foundation of basic physical understanding of semiconductors, which emerged from successful developments in other fields of physics. By quantum mechanical treatment of the microscopic crystal structure, the energy band structure can be derived which reflects the macroscopic properties of the semiconductor. From this band structure, carrier concentrations can be computed using the Fermi-Dirac distribution derived from statistical thermodynamics. These in turn, can be used to predict measurable quantities such as the current using transport physics. Despite many simplifications made along the way, this approach is capable of reproducing the output characteristics of many semiconductor devices with practical accuracy. Thereby making it a powerful tool for the understanding, development and design of semiconductor devices.

In this chapter, the theoretical background will be covered that is necessary to understand this thesis. The first section serves as introductory theory and recapitulates the most important basic semiconductor device physics. In the section thereafter, attention will be devoted to LDMOS integrated power devices, which are the special class of devices that are of interest in this research. The last two sections elaborate on ways that are used in this work to improve the traditional trade-off for integrated power devices.

2.1 Basic semiconductor device physics

One of the biggest advantages of using semiconductors for devices, is that the electrical characteristics can easily be altered by the use of doping (n- or p-type). Moreover, when a junction between a p-type and a n-type semiconductor region is formed, interesting physics arises which is current rectifying in nature. Generally, most semiconductor devices contain at least one such a pn-junction, such that the device characteristics and operation are inherently connected to it. In addition, also most of the terminology and concepts that are used in the analysis of devices also appear in the discussion of the pn-junction. Therefore, before going into more complex devices, some of the most important physics and results of the pn-junction will be briefly recapitulated here.

When a uniformly doped p-type material (acceptor doping concentration N_a [cm^{-3}]) forms a junction with a uniformly doped n-type material (donor doping concentration N_d [cm^{-3}]), a very large concentration gradient at the junction for both holes and electrons will exist. Therefore, majority electrons in the n-region will diffuse to the p-region and majority holes from the p-region will diffuse to the n-region. As a result, positive donor charge is left behind in the n-type material and negative acceptor charge in the p-type material (Figure 2). This region is referred to as space charge region or depletion region (ranging from $-x_p$ [cm] to x_n [cm]), no mobile charges are present here. The space charge within the depletion region sets up an electric field pointing from the n- to p-region (Figure 2). Consequently, there will be a potential drop along the depletion region which is called the built-in voltage (V_{bi} [V]). This V_{bi} produces a drift force ($\propto -\nabla V$) that works in the opposite direction as the diffusion force ($\propto \nabla n$ and $\propto \nabla p$) resulting from the carrier concentration gradient at each edge of the depletion region. In thermal equilibrium, these forces exactly cancel each other (Fermi level constant through the system). [11]

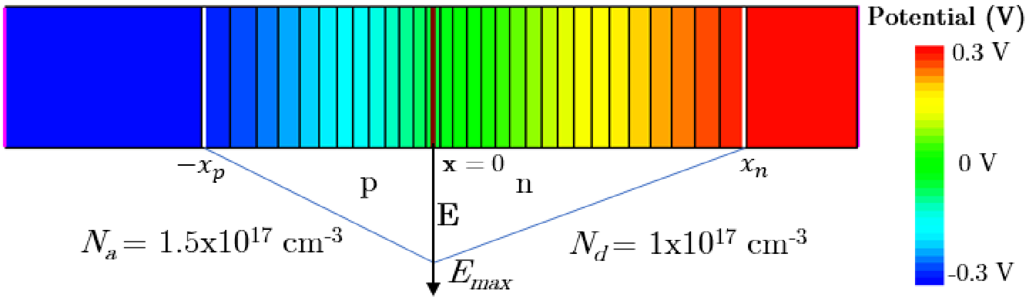

Figure 2: Simulation of a pn-junction at zero bias (V_R=0V) where N_a=1.5x10^{17} cm^{-3} and N_d=1x10^{17} cm^{-3}. Here the color indicates the magnitude of the electrostatic potential, the black lines the equipotential lines, the white lines the depletion boundaries and the brown line the junction boundary between the n-type and p-type material. Also, the electric field distribution is illustrated.

Without rigorous derivation, the most important electrostatic characteristics of an ideal pn-junction with an arbitrary bias (V_R [V]) applied to the n-region are given below: [11]

$$V_{bi} = \frac{k_B T}{e} \ln\left(\frac{N_a N_d}{n_i^2}\right) \tag{a}$$

$$x_n = \left(\frac{2\epsilon_s (V_{bi}+V_R)}{e}\left(\frac{N_a}{N_d}\right)\left(\frac{1}{N_a+N_d}\right)\right)^{\frac{1}{2}} \quad \& \quad x_p = \left(\frac{2\epsilon_s (V_{bi}+V_R)}{e}\left(\frac{N_d}{N_a}\right)\left(\frac{1}{N_a+N_d}\right)\right)^{\frac{1}{2}} \tag{b}$$

(1)

$$E(-x_p \le x \le 0) = \frac{-eN_a}{\epsilon_s}(x+x_p) \quad \& \quad E(0 \le x \le x_n) = \frac{-eN_d}{\epsilon_s}(x_n - x) \tag{c}$$

$$E_{max} = -\frac{2(V_{bi}+V_R)}{x_n + x_p} \tag{d}$$

Here $E(x)$ is the electric field distribution [V/cm], E_{max} the maximum electric field [V/cm], ϵ_s the dielectric permittivity (for Si: 11.7ϵ_0, where ϵ_0=8.85x10^{-14} F/cm), e the elementary charge (1.6x10^{-19} C),

T the temperature [K], k_B the Boltzmann constant (1.38x10^{-23} J/K) and n_i the intrinsic carrier concentration (for Si: 1.5x10^{10} cm^{-3}).

In contrast to Figure 2, in semiconductor devices there will always be a bias applied to the pn-junction such that it is no longer in thermal equilibrium (Fermi level not constant through the system). When V_R is positive (reverse bias configuration), the total potential barrier becomes larger than the zero-bias case. Moreover, since the electric field originates on positive charge and terminates at negative charge, the number of positive and negative charges in the depletion region must increase. As a result, the space charge width (x_n+x_p) increases with applied reverse bias voltage (Equation (1b)). The increased potential barrier withholds the majority electrons in the n-region to diffuse into the p-region and the majority holes in the p-region to diffuse into the n-region, such that there is essentially no (significant) current. When V_R is negative (forward bias configuration), the total potential barrier is lowered. By the same reasoning as before, the depletion width now decreases with increasing applied forward bias voltage. Due to the decreased electric field (or potential barrier) in the depletion region, the electrons in the n-region and holes in the p-region are now no longer held back. Hence, there will be a diffusion current of electrons from the n- to p-region and of holes from the p- to n-region. [11]

So, current can flow through the pn-junction when forward biased but is blocked when reverse biased. This characteristic can easily be used to make an electrical switch (i.e. a transistor) by combining two pn-junctions. The most used transistor in modern IC technology is the MOSFET and will be discussed next.

2.1.1 MOSFET

In general, a MOSFET consists of a source (S), drain (D), gate (G), body (B) and insulating spacer layer below the gate (Figure 3 (left)). The source and drain terminals are both highly doped silicon (n$^+$ in n-type and p$^+$ in p-type). The voltage across the channel is applied to the drain such that the current flow corresponds with the carriers (electrons in n-type, holes in p-type) flowing from the source to the drain. The body is respectively p/n-type doped silicon for n/p-channel devices. The gate is a highly conducting material (highly doped polysilicon) on top of an insulating material (generally SiO_2). In the ideal case, there is no current through the oxide to the gate terminal. In the rest of this section, only the n-channel MOSFET will be considered. For the p-channel MOSFET, the same arguments and relations apply only with holes being the charge carrier and with reversed current directions and voltage polarities.

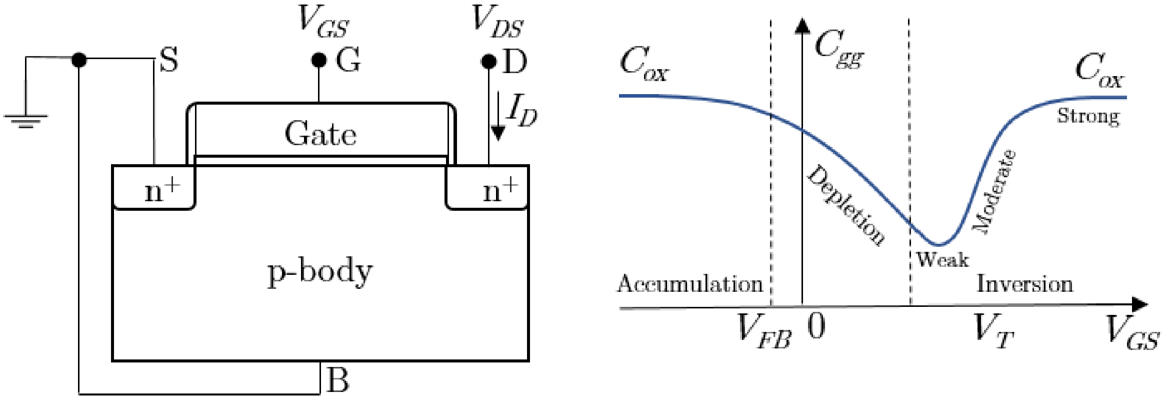

Figure 3: Left: Schematic representation of a n-channel MOSFET (S=Source, G=Gate, D=Drain and B=Body). Right: Typical low frequency gate capacitance (C_{gg}) characteristic of a n-channel MOSFET. Here also the accumulation, depletion and inversion regimes are indicated. [11]

Switching of the device is controlled by the voltage on the gate. The Metal-Oxide-Semiconductor (MOS) structure basically acts as a parallel plate capacitor, such that understanding its behavior can best be understood by looking at the gate capacitance (C_{gg}) as a function of the gate voltage (V_{GS}) (Figure 3 (right)). Depending on the sign of V_{GS}, the carriers can either be attracted towards the Si-SiO_2 interface (accumulation) or pushed away from it (depletion) resulting in a space charge region. When V_{GS} becomes higher than a certain threshold voltage (V_T [V]), the minority carrier concentration (electrons) at the Si-SiO$_2$ interface becomes higher than the majority carrier concentration (holes) in the bulk, an inversion layer is said to be formed. This inversion layer then "connects" the source and drain regions, such that a current can flow (I_D [A]). A MOSFET is therefore said to be in the off-state when $V_{GS} < V_T$ and in the on-state when $V_{GS} > V_T$.

Clearly, the V_T is one of the important parameters of the MOSFET since it determines whether the device is in the on- or off-state. Therefore, it should always be within the voltage range of a circuit design. It can mathematically be derived by calculating the gate voltage at which the band bending at the Si-SiO$_2$ interface is such that the Fermi level (E_f) is at least as far above the intrinsic Fermi level ($E_{f,i}$) as E_f is below $E_{f,i}$ in the bulk. By considering a simplified electrostatic charge distribution, that is approximately valid for MOSFET structures with long and wide channels, it can be derived that V_T obeys: [11], [39]

$$V_T = \frac{eN_a x_{dT}}{C_{ox}} + V_{FB} + 2\phi_{fp}$$
$$\text{where } x_{dT} = \left(\frac{4\epsilon_s \phi_{fp}}{eN_a}\right)^{\frac{1}{2}}, \qquad V_{FB} = \phi_{ms} - \frac{Q'_{ss}}{C_{ox}} \qquad \text{and} \qquad \phi_{fp} = \frac{k_B T}{e}\ln(\frac{N_a}{n_i}) \tag{2}$$

Here ϕ_{fp} ([J]) indicates the difference between the Fermi level (E_f) and the intrinsic Fermi level ($E_{f,i}$) in bulk. The x_{dT} ([cm]) represents the maximum space charge width at $V_{GS} = V_T$, a higher V_{GS} only increases the inversion charge density at the interface which shields the electric field from further penetrating the silicon. The flat-band voltage V_{FB} [V], is defined as the applied gate voltage at which there is no band bending in the silicon (Figure 3 (right)). The capacitance of the MOS capacitor in the strong inversion regime is denoted by C_{ox} (Figure 3 (right)), which for a simple parallel plate capacitor is given by t_{ox}/ϵ_{ox}, where t_{ox} ([cm]) and ϵ_{ox} ([F/cm]) are the oxide thickness and dielectric permittivity respectively. The work function difference (ϕ_{ms} ([J])) must be incorporated since the energy difference between the Fermi energy and the vacuum level will be different for silicon than for the gate material. The Q'_{ss} term ([C/cm^2]) represents the effective charge per unit area in the oxide at the Si-SiO$_2$ interface. This oxide charge is extremely unwanted in MOSFET's, but almost inevitable. Already during fabrication this oxide charge exists by e.g. structural defects (dangling bonds) and ionized impurities [45]. During the lifetime of the device, this oxide charge gets even built up more by e.g. ionizing radiation and injection and trapping of hot (high energetic) carriers in the oxide (HCI) [45], [46]. Therefore, the oxide charge can give rise to large unpredictable shifts and instabilities in V_T. Moreover, if this charge is closely situated near the Si-SiO$_2$ interface, the mobility is degraded thereby altering the current characteristics. [11], [39]

2.1.2 Current characteristics of a MOSFET

The basic operation of a MOSFET is modulating the channel resistance by the applied gate voltage. The resistance in the channel depends on the dimensions, the mobility and the inversion charge present. In the on-state for $V_{DS} < V_{GS} - V_T$, the following expression then applies for the channel resistance R_{ch} [Ω]: [11], [14], [39]

$$R_{ch} = \frac{L_{ch}}{W\mu_{ch}|Q'_{inv}|} = \frac{L_{ch}}{W\mu_{ch}C_{ox}(V_{GS} - V_T - V_{DS}/2)} \tag{3}$$

Here L_{ch} is the length of the channel and W the width of the device, both in [cm]. $|Q'_{inv}|$ is the inversion layer charge per unit area made up from electrons [C/cm^2]. This charge is a function of the gate voltage and is given by $C_{ox}(V_{GS} - V_T - V_{DS}/2)$ in the strong inversion regime. μ_{ch} is the effective electron mobility in the inversion channel [cm^{2}V^{-1}s^{-1}]. This mobility is typically a factor two lower than in bulk, because of additional surface scattering due to the gate induced electric field which accelerates the inversion carriers to the surface [14]. In section 2.4, more attention will be devoted to the channel mobility.

So, when a small drain voltage is applied ($V_{DS} \to 0$) and the gate voltage is increased, the inversion charge density starts to increase linearly from $V_T + V_{DS}/2$ (Equation (3)) such that also I_D increases ideally linearly (Figure 4 (left)). This purely linear increase however, is not entirely true. Already below $V_T + V_{DS}/2$ there will be some current flow, which is referred to as subthreshold conduction. This is because before reaching the inversion point, the channel is in the weak/moderate inversion regime (Figure 3 (right)) where E_f is closer to the conduction band than to the valence band such that the channel has the characteristics of a lightly doped n-type semiconductor. So MOSFET's should always be biased sufficiently below the V_T in the off-state to minimize off-state power dissipation. Moreover, since the mobility degrades with V_{GS} (section 2.4), the slope of the $I_D - V_{GS}$ curve (i.e. the transconductance (g_t)) decreases (Figure 4 (left)). [11], [46]

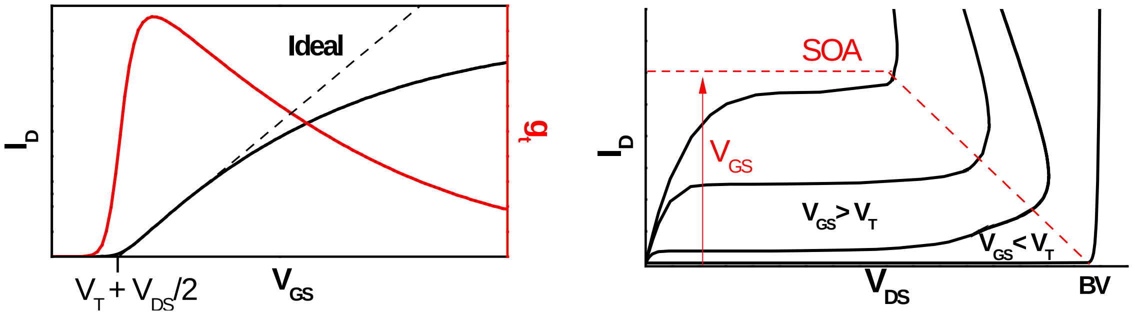

Figure 4: Left: Example of drain current and transconductance as a function of the gate voltage. Also, the ideal drain current is indicated. Right: Example of drain current as a function of the drain voltage for increasing gate voltage. Outside the Safe Operating Area, the current increases uncontrollably due to breakdown mechanisms.

When now V_{GS} is held constant and V_{DS} is increased, the typical curves shown in (Figure 4 (right)) are obtained. In the on-state, for $V_{DS} \to 0$, the channel acts as an Ohmic resistor (Equation (3)) where the initial slope (R_{ch}^{-1}) is determined by V_{GS}. As V_{DS} starts to increase, the voltage drop at the drain side across the oxide decreases, such that the inversion charge density decreases locally. As a result, the incremental resistance increases with increasing V_{DS} up to the point where the voltage drop across the oxide at the drain is equal to V_T, such that the inversion charge density becomes zero at the drain terminal (pinch-off). If V_{DS} increases further, the point where the inversion charge becomes zero moves to the source. At this point electrons are injected into the space charge region and subsequently drift to the drain. For large channels, the drain current then will be constant after the pinch-off. Therefore, the MOSFET is said to be operating in the saturation regime for drain voltages larger than $V_{DS}(sat) = V_{GS} - V_T$. In this regime, as well as in the linear regime, the ideal current-voltage relationships follow from the square-law model and are given by: [11], [14], [39]

$$\begin{aligned} V_{DS} < V_{DS}(sat):\quad I_D &= \frac{W\mu_{ch}C_{ox}}{2L_{ch}}(2(V_{GS}-V_T)V_{DS} - V_{DS}^2) \\ V_{DS} > V_{DS}(sat):\quad I_D &= \frac{W\mu_{ch}C_{ox}}{2L_{ch}}(V_{GS}-V_T)^2 \end{aligned} \tag{4}$$

It must be noted that these equations do not include non-idealities such as: subthreshold conduction, channel length modulation, mobility variation and velocity saturation. These and other non-idealities however, become only significant for short channel devices [11]. Power devices typically have relatively long channels such that these non-idealities have little influence. In addition, from (Figure 4 (right)) it can also be seen that these equations only hold in the so-called Safe Operating Area (SOA). Outside this regime, the current starts to increase uncontrollably and the device is said to be in breakdown.

2.1.3 Breakdown in a MOSFET

In a MOSFET, several current generation mechanisms must be considered that can make the device go into breakdown. They all have in common that they start occurring when high electric fields are present. The electric fields are generally the highest at pn-junctions and at semiconductor-oxide interfaces, which are therefore the most profound locations for breakdown to set in.

The most destructive physical mechanism, is impact ionization occurring in depleted regions with high electric fields (such as in the reverse biased pn-junction at the drain side). This is the process in which electron-hole pairs are generated by impact of silicon atoms by hot carriers. This is particularly harmful because the generated charge carriers in turn can ionize other silicon atoms resulting in charge carrier multiplication. The efficiency of this process (in 1D) can be expressed by the charge multiplication factor (M): [11], [47]

$$M = \int_0^{W_d} \alpha(E)dx \tag{5}$$

Here W_d is the depletion width [cm] and α the reciprocal of the mean free path between two impact ionizations also called the ionization coefficient [cm^{-1}]. Note that here it is assumed that the ionization coefficient for electrons is the same as for holes (i.e. $\alpha_n = \alpha_p = \alpha$). This coefficient is a strong exponential function of the electric field as dictated by Chynoweth's law [14]. For analytic derivations however, Chynoweth's law is often approximated by Fulop's power law for silicon: $\alpha = 1.8 \cdot 10^{-35} E^7$ [14].

The charge multiplication process, depending on M, can result in two destructive breakdown mechanisms. The typical breakdown mechanism considered occurs for $M \geq 1$ and is called *avalanche breakdown*. In avalanche breakdown, each charge carrier at least ionizes one other atom thereby resulting in a self-amplifying (uncontrollable) current. For $M < 1$ on the other hand, before avalanche breakdown sets in, *snapback breakdown* may already make the device uncontrollable. This is a second order effect caused by the parasitic bipolar (npn) formed by the source-body-drain structure. While the generated electrons flow to the drain, the holes flow to the body terminal resulting in a voltage difference between the source and body due to the resistance of the body. This may forward bias the source-body junction resulting in electron injection in the body. Part of these injected electrons then diffuse to the drain resulting in a positive feedback loop. This type of breakdown can be recognized by a negative slope in the $I_D - V_{DS}$ curve, as can be seen in Figure 4 (right). Both the impact ionization based breakdown mechanisms are considered as abrupt on the $I_D - V_{DS}$ curve due to the self-amplifying nature. [11], [47]

Next to the previously discussed abrupt breakdown mechanisms, there are also more subtle mechanisms that are typically less destructive. One of them is *(near) punch-through breakdown*. Punch-through is the effect when the depletion region of the drain junction fully extends to the depletion region of the source junction. In that case, the energy barrier between the source and drain is completely gone resulting in a large drain current. Although, also already before punch-through is reached, the drain current rapidly starts to increase. This is due to a phenomenon called drain induced barrier lowering (DIBL), which is the lowering of the energy barrier due to the interaction of the source and drain depletion regions. Since the drain current depends exponential on the barrier height, the current increases rapidly with V_{DS} when the near punch-through condition is reached. This breakdown mechanism can be suppressed by using extra doping in the body near the source and the drain, which are also called halo implants. [11], [39]

Another subtle breakdown mechanism is *Gate Induced Drain Leakage (GIDL)*. This drain leakage mechanism is due to quantum mechanical band-to-band tunneling. In the presence of high electric fields (~ >80 V/µm), such as in heavily doped pn-junctions or MOS structures, phonon-assisted band-to-band tunneling generally cannot be neglected. GIDL in particular, relates to the band-to-band tunneling near the Si-SiO$_2$ interface below the gate which becomes significant when there is a large voltage difference between the drain and the gate. This leakage mechanism is typically the most dominant one before impact ionization based breakdown mechanisms set in. [47], [48] An elegant approach to reduce breakdown effects, without affecting the device's geometry or dimensions, are LDD (Lightly Doped Drain) implants [11], [39]. In this approach, lightly doped regions are situated next to the drain and the source of the same type of doping. The reduced doping gradient in going from the channel to the source/drain lowers the (peak) electric field and thereby reduces breakdown effects. This type of concept is actually also used for (LDMOS) power devices, in which the drain is extended by a lower doped region to achieve high breakdown voltages.

2.2 LDMOS integrated power devices

Integrated power devices need to handle large currents in the on-state and reach high (off-state) breakdown voltages. One way to accomplish this, would be by simply upscaling the standard MOSFET structure. This however, is beyond certain voltages not a cost-efficient and straightforward way to increase the voltage and current handling capabilities. For example, if the off-state breakdown voltage and the device dimensions are to be scaled with a factor k (>1), the doping in the substrate should be scaled with k^{-1} (<1) [11]. This leads to large modifications in V_T (Equation (2)). Moreover, the mobility in the inversion channel is relatively low compared to the bulk mobility [14], such that an even longer inversion channel would not be efficient. Therefore, generally the way to go for power devices is a drain-extension. This is a lowly doped region connecting the channel with the drain, with respectively n/p-type doping for a n/p-channel MOSFET. For integrated applications, such a device is typically called a LDMOS (laterally diffused MOS) and is shown in the left figure below. It should be mentioned that this name convention originally refers to devices in which the channel is implemented by lateral diffusion, however is nowadays generally also used to refer to drain-extended devices fabricated by other techniques as well [49].

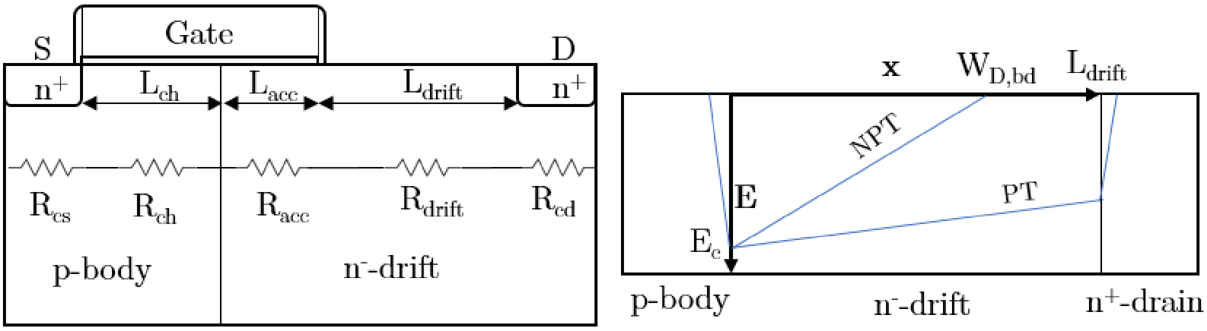

Figure 5: Left: Schematic illustration of a LDMOS power device, which has a lowly doped drain-extension that connects the channel to the drain. Here also the different series resistance contributions are indicated.
Right: Schematic illustration for a simplified 1D off-state breakdown analysis of the LDMOS device (in which L_{acc} can be set to zero) for both the NPT (non-punch through) condition and the PT (punch through) condition.

Just as in the normal MOSFET, for sufficient gate bias an inversion layer forms in the p-body below the gate in which electrons are injected for $V_{DS} > 0$. Only now after passing through the inversion layer, the electrons flow into the drain-extension where they are accelerated to the drain by the electric field. Commonly, there is also some gate overlap on the drain-extension (L_{acc}), as depicted in Figure 5 (left). In the on-state, this enhances the population of charge carriers near the surface (accumulation), which can be beneficial for the resistance. In the off-state, this extended gate works as a field plate which optimizes the electric field in the drain-extension and thereby increases the off-state breakdown voltage (subsection 2.3.1 will go more into detail about field plates). [1]

In order to handle large currents in the on-state and to block large drain-to-source voltages in the off-state, both the on-state resistance (R_{ON}) and the off-state breakdown voltage (BV) must be optimized as much as possible without compromising the other. This is therefore the main task in designing power devices (for switching applications). Hence, in the rest of this section these two key design targets will be discussed in more detail for the LDMOS power device.

2.2.1 On-state resistance

The R_{ON} of the LDMOS can be decomposed in several contributing resistances, as shown in Figure 5 (left). These resistances are in series such that for R$_{ON}$ [Ω] in the linear regime ($V_{DS} < V_{DS}(sat)$), the following equation can be written: [50]

$$R_{ON} = \frac{\partial V_{DS}}{\partial I_d} = R_{cs} + R_{ch} + R_{acc} + R_{drift} + R_{cd} \tag{6}$$

The typical drain bias that is used in NXP's technology platform to measure the R$_{ON}$ in this regime is 0.1V at full gate drive. The advance of measuring (and modeling) in this regime is that non-idealities caused by significant drain bias are limited [11], [46]. For example, for a small channel length of 0.2 μm the lateral field is 0.5 V/μm, which is significantly below the 3 V/μm at which the electron velocity starts to saturate [46]. The limitation of this and other non-idealities makes it relatively easy to obtain expressions for the individual resistance contributions:

Source and drain resistance ($R_{cs} + R_{cd} = R_{sd}$ [Ω])
These resistances consist of the metal-silicon contact resistances (which depend on the work function of the metal and the doping concentration at the surface) and of the resistances in the highly doped source/drain regions itself [14]. These resistances are relatively small and are only weakly dependent on the terminal voltages [46], such that for simplicity they are assumed to be constants.

Channel resistance (R_{ch} [Ω])
The channel resistance (R_{ch}) for an intrinsic MOSFET was already given by Equation (3). However, the correction for the average channel potential ($-V_{DS}/2$) is not valid for a LDMOS device since there is an additional voltage drop over the drain-extension. The resistance of the drain-extension (R_{acc}+R_{drift}) is typically must larger than R_{ch}, such that the average channel potential safely can be omitted for a low drain bias (such as V_{DS}=0.1V). The channel resistance is then given by: [14], [50]

$$R_{ch} = \frac{L_{ch}}{W\mu_{ch}|Q'_{inv}|} = \frac{L_{ch}}{W\mu_{ch}C_{ox}(V_{GS} - V_T)} \tag{7}$$

Accumulation resistance (R_{acc} [Ω])
The accumulation resistance (R_{acc}) is the resistance in the silicon area where the gate or field plate overlaps the drain-extension and in literature is most commonly expressed as: [14], [50]

$$R_{acc} = \frac{L_{acc}K_{acc}}{W\mu_{acc}|Q'_{acc}|} = \frac{L_{acc}K_{acc}}{W\mu_{acc}C_{ox}(V_{GS} - V_T)} \tag{8}$$

Here K_{acc} is a factor (<1) to account for current spreading [-], L_{acc} the length of the accumulation region [cm], $|Q'_{acc}|$ the accumulation charge per unit area [C/cm^2] and μ_{acc} the mobility in the accumulation layer [cm^{2}V^{-1}s^{-1}]. The mobility in the accumulation layer is generally higher than in the inversion layer, since the charge carriers are distributed further from the interface than in the inversion layer [14], thereby reducing surface roughness scattering. It has been reported to be ~80% of the bulk mobility [14].
Although Equation (8) describes an inverse V_{GS} dependence as is expected for accumulated charge, one consideration on which this equation is based may not be fully valid in this work. Namely, it is assumed that the accumulation charge is built up from the channel threshold voltage. Obviously, this is theoretically not true because the charge is accumulated from the flat-band voltage (V_{FB}) of the accumulation layer [51], [52]. So, unless $V_{T,channel} \approx V_{FB,accumulation\ layer}$, this equation cannot be assumed to be trustworthy. In Appendix A.1, it is shown how V_{FB} can be calculated and what values can be expected for the

accumulation regions in this work. For a typical donor doping concentrations of 10^{17} cm^{-3}, V_{FB} is approximately -0.15V [14]. Typical threshold voltages are in the range from 0.5 - 1.5V, so clearly the condition does not apply. The alternate proposal for the accumulation resistance would then be:

$$R^*_{acc} = \frac{L_{acc}K_{acc}}{W\mu_{acc}C_{ox}(V_{GS} - V_{FB})} \tag{9}$$

One final thing that should be mentioned about Equations (8) and (9), is that it is assumed that the accumulation region is not biased (or equivalently $V_{DS} \to 0$). Taking this into consideration, would effectively increase V_{FB} by the average potential in the accumulation region ($\overline{V_{acc}}$). For a low drain bias however (such as V_{DS}=0.1V), $\overline{V_{acc}}$ is typically negligible because most of the voltage is already supported over the drift region.

Drift resistance (R_{drift} [Ω])
The resistance of the drift region is just that of a doped semiconductor sheet and is typically expressed as: [50]

$$R_{drift} = \frac{L_{drift}}{eN_d\mu_n d_{eff}W} \tag{10}$$

Here d_{eff} is the effective well thickness in the drift region to account for the unequal distribution of the current flow in the drift region (current spreading) [cm], L_{drift} the length of the drift region [cm], N_d the donor concentration in the drift region [cm^{-3}] and μ_n the bulk mobility appropriate to N_d [cm^{2}V^{-1}s^{-1}]. This equation implies that R_{drift} is constant and independent of any of the terminal voltages, which is particularly true in the limit that V_{DS} →0. If however this limit is not valid, R_{drift} is typically related to V_{DS} and V_{GS} by phenomena such as the Kirk effect, JFET resistance (subsection 2.3.2), self-heating and quasi saturation (velocity saturation in drift region) [53]–[55]. For this work, it is assumed that V_{DS}=0.1V is sufficient low for this simplification to be usable.

Total resistance (R_{ON} [Ω])
Filling in all the resistance contributions (Equations (7), (9) and (10)) in Equation (6) and setting R_{cs}+R_{cd}=R_{sd}, gives the following expression for R_{ON} valid for low drain bias:

$$R_{ON} = \frac{L_{ch}}{W\mu_{ch}C_{ox}(V_{GS} - V_T)} + \frac{L_{acc}K_{acc}}{W\mu_{acc}C_{ox}(V_{GS} - V_{FB})} + \frac{L_{drift}}{eN_d\mu_n d_{eff}W} + R_{sd} \tag{11}$$

It must be mentioned that in this equation the mobility reduction with increasing V_{GS} in inversed/accumulated layers is not accounted for. These mobility reductions are typically modeled as: [46]

$$\mu_{ch} = \frac{\mu_{0,ch}}{1 + \theta_{ch}(V_{GS} - V_T)} \quad \& \quad \mu_{acc} = \frac{\mu_{0,acc}}{1 + \theta_{acc}(V_{GS} - V_{FB})} \tag{12}$$

Here $\mu_{0,ch}$ and $\mu_{0,acc}$ are the zero-field bulk mobilities in the channel and accumulation region respectively [cm^{2}V^{-1}s^{-1}] and θ_{ch} and θ_{acc} are technology defined constants [-]. Including these in Equation (11) would result in two additional (constant) series resistances proportional to the θ constants, which have the same effect on R_{ON} (V_{GS}) as R_{sd} and R_{drift}.

To give an impression of the influence of the various resistance contributions on the R_{ON} and on the linear drain current as a function of V_{GS}, Equation (11) is used to generate an example which is shown in Figure 6. Here R_{res} denotes the constant resistance contributions. From Figure 6, it can be seen that after the current rises rapidly due to the fast decrease in R_{ch} and R_{acc}, its starts to saturate because R_{res} becomes dominant. Typically, especially for large drift regions, R_{drift} will be the largest contribution to R_{res}. Therefore, for proper modeling of LDMOS devices, incorporation of the mobility reduction with increasing V_{GS} is not as crucial as for standard MOSFET's.

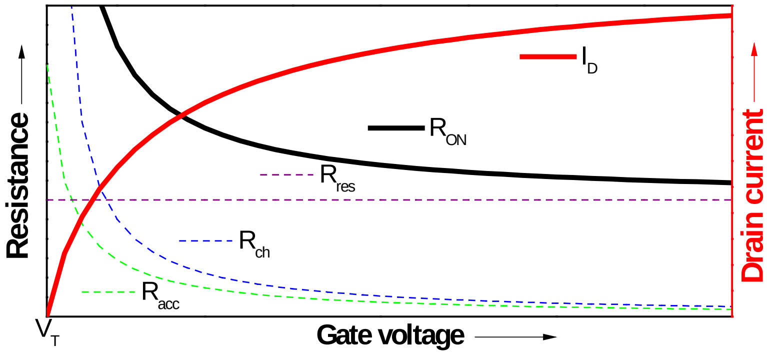

Figure 6: Typical example of the R_{ON}, as well as its contributions, and the linear drain current as a function of V_{GS} for a n-type LDMOS device. The current starts rising rapidly due to the fast decrease in R_{ch} and R_{acc} and eventually saturates because R_{res} becomes dominant. Here R_{res} denotes the constant resistance contributions (R_{sd} and R_{drift}). In addition, also the mobility reduction in inversed/accumulated layers can be incorporated in R_{res}, although can typically be neglected with respect to R_{drift}.

It must again be noted that Figure 6 is just an example. The distribution of the various contributions to the total resistance very depends on the application voltage. For high breakdown voltages (>25V), the drift region will be relatively long such that R_{drift} will be dominant [50]. For lower breakdown voltages however, the drift region will be short such that R_{ch}, R_{acc} and R_{drift} may be on the same order [50]. The relative contributions to the resistance are particularly of interest for identifying possibilities for optimization. In Appendix A.2, a method is proposed to separate the different resistance contributions in a LDMOS device based on the theory in this subsection.

To compare the total on-state resistance of power devices, it is convenient to use the specific on-state resistance (R_{ON}A $[m\Omega \cdot mm^2]$), which was already stated to be an important FOM in section 1.1. This is just the R_{ON} multiplied by its surface area (A), in which the total length of the device (pitch) is the distance from the middle of the body contact to the middle of the drain contact (since the body and drain contacts are shared). The other important FOM, namely the BV, will be discussed next.

2.2.2 Off-state breakdown voltage

When the LDMOS device is in the off-state and a positive drain voltage is applied, the junction between the body and the drain-extension is reverse biased. The doping of the drain-extension is typically lower than that of the channel region, such that the junction is usually assumed to be one-sided (i.e. p^+n or pn^-). As a result, the depletion region mainly extends into the drain-extension (Equation (1b)), such that most of the applied voltage is supported there. Depending on the length of the drain-extension and its doping, two situations can then be considered. The non-punch through situation (NPT), in which the electric field at the drain edge remains zero. And the punch-through condition (PT), in which the depletion region fully extends to the drain such that the electric field at the drain edge is non-zero. In Figure 5 (right), these situations are sketched for a simplified 1D off-state breakdown analysis. Note that L_{acc} is set to zero here as the influence of the gate can be neglected in one dimension. Most devices are generally designed to operate in the PT condition, since this yields better R_{ON}A-BV trade-off (as will be shown in Appendix B.2) [1]. However, breakdown related relations are typically only derived for the NPT case since this results in much more manageable expressions. Moreover, these are independent of L_{drift}, such that a closed form equation can be obtained for the R_{ON}A-BV trade-off. [1], [14], [56] For the NPT case, using Equation (1c) (in the limit that $N_d \to 0$) together with Fulop's approximation, it can then be derived by using Equation (5) (with M=1) that the BV due to avalanche breakdown and corresponding depletion width ($W_{D,bd}$) for silicon are given by: (derivation in Appendix B.1) [14]

$$\begin{aligned} BV &= 5.34 \cdot 10^{13} \cdot N_d^{-\frac{3}{4}} && \text{a} \\ W_{D,bd} &= 2.67 \cdot 10^{10} \cdot N_d^{-\frac{7}{8}} && \text{b} \end{aligned} \tag{13}$$

Here N_d denotes the donor concentration of the lowly doped drift region [cm^{-3}]. Note that this BV only applies for a one-sided planar junction. In practice, the junction between the drain-extension and the body can be curved which enhances the electric field at the edges. This limits the breakdown to about 80% of the planar junction [14].

Generally, the voltage drop over the drift region can be obtained by integrating the electric field over the drift region:

$$V_{drift} = -\int_0^{L_{drift}} \vec{\boldsymbol{E}} \cdot \vec{\boldsymbol{dl}} = -\int_0^{L_{drift}} E_x(x,y)dx \tag{14}$$

At breakdown conditions (situation sketched in Figure 5 (right)), this then gives: $BV = E_c W_{D,bd}/2$. Using the equations in (13), the following expression can then be derived for the critical field: [14]

$$E_c = 4010 \cdot N_d^{\frac{1}{8}} \tag{15}$$

The critical electric field (E_c) is a particular useful parameter for identifying the onset of avalanche breakdown. This is because the impact ionization coefficient (α) depends strongly on the electric field [14]. Avalanche breakdown can therefore usually be assumed to occur whenever the electric field approaches the critical electric field locally at some point in the device. As a rule of thumb, E_c is typically assumed to be on the order of 20 V/µm for silicon [9], [10].

Too check how the solutions of the NPT case compare to the solutions of the PT case, the BV and E_c are plotted in the figure below for both cases for five different drift lengths. For the BV of the PT case the expression in reference [56] is used, this is also used to derive E_c in a similar manner as Equation (15) was derived.

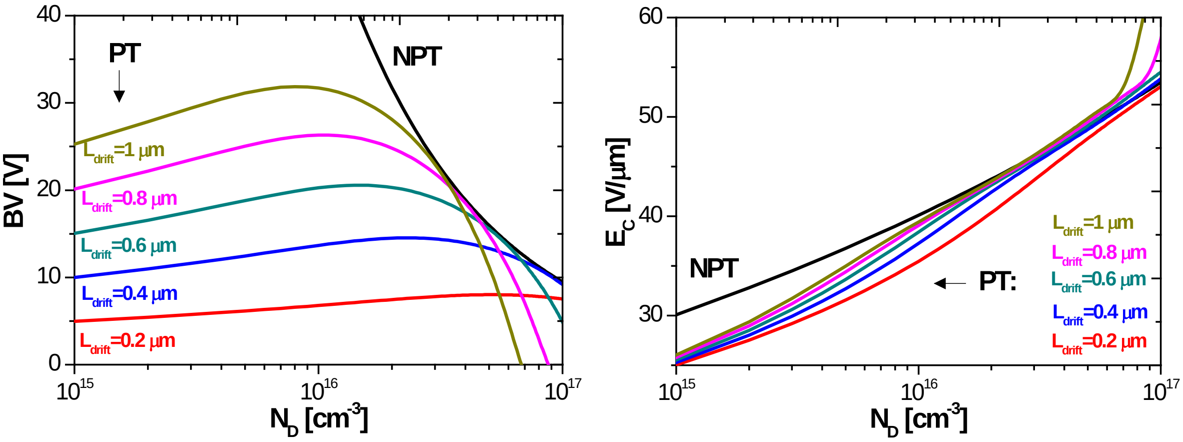

Figure 7: BV (left) and E_c (right) for the NPT- and PT case for five different drift lengths. The PT solutions converge to the NPT solutions for increasing doping concentration and length of the drift region. Note that when the NPT and PT curves cross, $W_{D,bd} = L_{drift}$, such that the transition from PT to NPT occurs.

From Figure 7, it can be seen that with increased doping concentration and length of the lightly doped drift region, the BV and E_c of the PT case converge to the NPT case. Note that when the NPT and PT curves cross, $W_{D,bd} = L_{drift}$, such that the transition from PT to NPT occurs. In this work, typically L_{drift} ≈0.4 µm and N_d ≈10^{17} cm^{-3}, for which the PT and NPT solutions are very close together such that the less complex NPT solutions safely can be used in further analysis.

Now that the two most important FOM's (i.e. R$_{\text{ON}}$A and BV) for (LDMOS) power devices are discussed in more detail, it will be explained why it is hard to optimize one without comprising the other

2.2.3 Traditional trade-off

The compromise between the BV and R$_{\text{ON}}$A is often referred to as the traditional trade-off in the power semiconductor device industry. In order to fully evaluate this trade-off, all components of the LDMOS that affect the field distribution, resistance and/or pitch should be taking into consideration. However, since the "intrinsic MOSFET" part (i.e. the channel region) does not support a large fraction of the off-state voltage, it can be approximated that it only adds a (small) offset to the R$_{\text{ON}}$A in the trade-off. On the other hand, the drain-extension supports most of the off-state voltage and is typically also the largest contributor to the R$_{\text{ON}}$A. Consequently, the trade-off between the BV and R$_{\text{ON}}$A finds its largest origin in the properties of the drain-extension and will therefore only be discussed for the drain-extension of the LDMOS. [1], [14]

In order to support a high BV over the drain-extension, it became evident from the previous subsection that the drift doping should be low enough (Equation (13a)). Moreover, since a low doping concentration results in a large depletion width (Equation (13b)), L_{drift} needs to scale accordingly. Unfortunately, both these requirements compromise the resistance of the drift region (Equation (10)). From a 1D analysis of an ideal LDMOS with a lowly uniformly doped drift region, the ideal trade-off between the specific on-state resistance of the drift region and the BV can be approximated by: (derivation in Appendix B.2)

$$R_{drift}A_{drift}\ [\Omega \cdot cm^2] = R_{drift} \cdot WL_{drift} = \frac{9.18 \cdot 10^{-15}}{d_{eff}} \cdot BV^{\frac{11}{3}} \tag{16}$$

It must be noted that Equation (16) will mostly be valid for LDMOS devices in the higher voltage range, since there the BV will approximately be fully supported by the drain-extension.

One way to break the ideal 1D trade-off limit, is the so-called RESURF concept. This concept enables to improve the BV without compromising the R$_{\text{ON}}$A and will be the topic of the next section.

2.3 Improving breakdown - RESURF

Using the Reduced SURFface Field (RESURF) effect [15], the drain-extension in LDMOS devices can be designed in such a way that the (off-state) breakdown voltage is maximized without sacrificing the on-state resistance. From subsection 2.2.2, it became evident that it can be assumed that breakdown occurs whenever the electric field becomes larger than the critical field (E_c) at a certain location in the silicon. Therefore, the maximal BV is obtained whenever the electric field is equal to E_c everywhere along the drain-extension. In other words, the BV is maximized by ensuring that the electric field along the current direction (x) in the drain-extension obeys $\frac{\partial E_x(x,y)}{\partial x} = 0$. This is the ideal RESURF condition.

In general, to design a drain-extension that obeys this condition, one has to mathematically describe the electric field distribution within the drain-extension. For an electrostatic system, the two relevant Maxwell equations are respectively Gauss's law and Faraday's law: [57]

$$\nabla \cdot \vec{\boldsymbol{E}} = \frac{\rho}{\epsilon_s} \quad \& \quad \nabla \times \vec{\boldsymbol{E}} = -\frac{\partial \vec{\boldsymbol{B}}}{\partial t} \tag{17}$$

Here ρ denotes the charge density [C/cm^{-3}], $\vec{\boldsymbol{E}}$ the electric field [V/cm], $\vec{\boldsymbol{B}}$ the magnetic field [Vs/cm^2] and t the time [s]. Since the system is assumed to be static, the time derivative is zero, meaning that the electric field is a conservative vector field such that it can be written as the gradient of a scalar potential: $\vec{\boldsymbol{E}} = -\nabla V(\vec{\boldsymbol{r}})$. Substituting this into Gauss's law, then yields Poisson's equation: [57]

$$\nabla^2 V = -\nabla \cdot \vec{\boldsymbol{E}} = -\frac{\rho}{\epsilon_s} \tag{18}$$

Here V is the electrostatic potential [V]. For a 1D drain-extension, $\frac{dE_x(x)}{dx} = \frac{\rho}{\epsilon_s}$, such that the RESURF condition cannot be met since $\rho \neq 0$. For a 2D drain-extension however, $\frac{\partial E_x(x,z)}{\partial x} + \frac{\partial E_z(x,z)}{\partial z} = \frac{\rho}{\epsilon_s}$, the ideal RESURF condition can be satisfied if there is a field present perpendicular to the current direction with gradient: $\frac{\partial E_z(x,z)}{\partial z} = \frac{\rho}{\epsilon_s}$. In Figure 8, both these cases are illustrated. For identical E_c and BV, an ideal 2D drain-extension only needs half of the length as compared to the 1D case. [10]

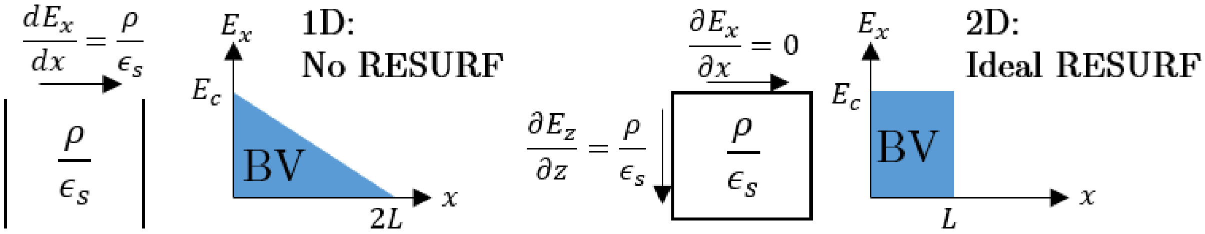
$\frac{dE_x}{dx} = \frac{\rho}{\epsilon_s}$
$\frac{\rho}{\epsilon_s}$
E_x
E_c
1D:
No RESURF
BV
2L
x
$\frac{\partial E_x}{\partial x} = 0$
$\frac{\partial E_z}{\partial z} = \frac{\rho}{\epsilon_s}$
$\frac{\rho}{\epsilon_s}$
E_x
E_c
2D:
Ideal RESURF
BV
L
x

Figure 8: Illustration of the RESRUF concept. For identical E_C and BV, a RESURF optimized (2D) drain-extension only needs half of the length compared to the 1D case. [15]

An additional advantage of using the RESURF concept, is that the drain-extension can be higher doped, thereby lowering R_{drift} (Equation (10)). While increasing the doping concentration would lead to a reduction of the BV for a 1D device (Equation (13a), RESURF can be used to counterbalance the increase in charge density by a higher perpendicular field gradient [10]. This combined with the optimized field distribution, reduces the traditional trade-off to: (derivation and full expression in Appendix B.3)

$$R_{drift}A_{drift}\ [\Omega \cdot cm^2] = \frac{1.64 \cdot 10^7}{d_{eff}N_d}\left(\frac{1}{\mu_n(N_d)}\right) BV^{\frac{7}{3}} \tag{19}$$

From Equation (19), it is now evident that increasing the drain-extension doping is very beneficial for the trade-off. However, this must be tempered by the consideration that the mobility becomes lower at high doping concentrations. In contrast to the 1D trade-off (Equation (16)), this mobility reduction is now incorporated because in RESURF applications the drift doping is typically higher. In addition, it must be mentioned that the doping concentration cannot be increased indefinitely due to the associated increase in perpendicular field and decrease in perpendicular depletion of the drain-extension. The perpendicular depletion should be at least as big as the perpendicular layer thickness t [cm] for any x along the drain-extension. This criterium prevents premature p-body/n-drift junction breakdown by ensuring that the whole drain-extension is depleted such that ideal RESURF conditions can apply [9], [10]. Moreover, in order to prevent premature perpendicular breakdown, the full perpendicular depletion of the drain-extension should occur before the (perpendicular) critical field is reached. Using Equation (1c), these two criteria can be translated in a mathematical expression for the maximum doping dose in the drain-extension: [10], [14]

$$N_d(x) \cdot t(x) \leq \frac{\epsilon_s E_c(x)}{e} \tag{20}$$

The perpendicular critical field $E_c(x)$ should follow from taking the ionization integral along the layer thickness $t(x)$. In literature, typically a critical field of 20 V/µm is assumed, resulting in a maximum doping dose on the order of 10^{12} cm^{-2} for silicon [1], [10], [14], [15].

So, from Equation (19) (combined with the restriction from Equation (20)) it is clear that the application of RESURF results in a major improvement in the traditional trade-off for LDMOS devices. The natural question then arises how such a perpendicular field gradient can be produced that is also compatible with the device structure. The two typical ways this is done will be discussed next.

2.3.1 Field plate assisted RESURF

The most intuitive way to generate a perpendicular field gradient is by adding a region/boundary on top of the drain-extension, that is biased by a certain voltage V_{FP}. Such a structure is generally called a field plate (FP) and can take many different forms. The part of the gate overlapping the drain-extension (i.e. an extended gate) in Figure 5 (left) is for example called a FP. But it could for instance also be a metal contact or a second gate like structure above the drift region. Generally, the FP is connected to the gate, such that in the blocking state V_{FP}=0V and allowing for additional accumulation in the on-state. In order to illustrate how a FP optimizes the electric field distribution in the drain-extension (in the off-state), a simulation is performed of a simplified p-body/n-drift/n-drain structure with and without a (metal) FP on top of the drain-extension separated by an oxide:

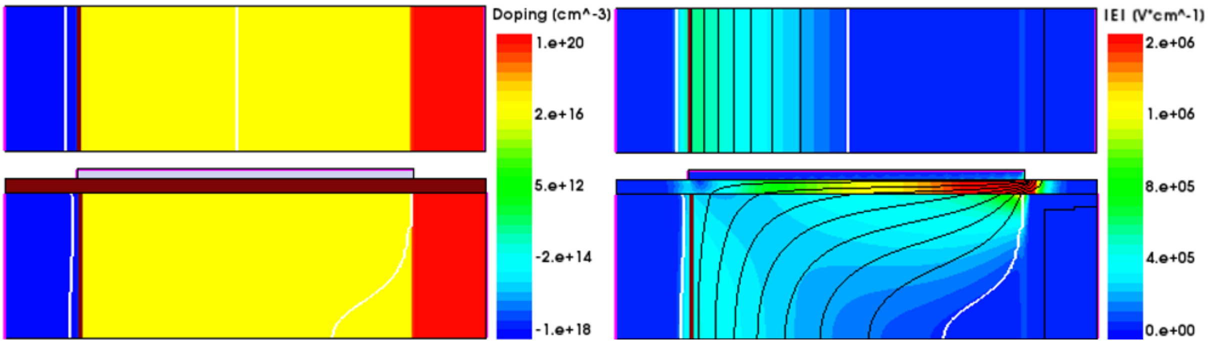

Figure 9: Simulation of a simplified p-body/n-drift/n-drain structure with (bottom) and without (top) a metal FP on top of the drain-extension separated by an oxide. The body doping is 10^{18} cm^{-3}, the drift doping 10^{17} cm^{-3} and the drain doping 10^{20} cm^{-3}. The applied bias at the drain (right contact) is 10V and the FP is at 0V. Left: Doping distribution. Right: Magnitude of electric field and equipotential lines. Note that the white lines indicate the depletion boundaries and the brown lines the junction boundaries.

From Figure 9, it can be seen that the metal FP keeps the electric field in the current direction more uniform by extending the depletion region (white line) and pushing the equipotential lines towards the drain (right contact). Countercharge for the depleted drain-extension is now not only present in the body, but also on the FP. This results in a vertical field gradient which thereby limits the horizontal field gradient. It must however be noted, that this field distribution is still far from ideal RESURF. The electric field now peaks at the right FP edge, which is therefore the most profound location for breakdown to set in. This is because the bias between the FP and the drain-extension increases towards the drain. Consequently, the FP will help deplete the drain-extension more towards the drain than towards the body. The vertical field gradient is thus not uniform over the drain-extension, such that ideal RESURF cannot apply. By solving the Poisson equation in the drain-extension at breakdown for ideal RESURF conditions and imposing the boundary condition: $V(x, \pm t_{eq}) = V_{FP}(x)$, the following condition for ideal RESURF can be obtained: [9], [10]

$$\frac{\epsilon_s\left(\frac{BV}{L_{DE}}x - V_{FP}(x)\right)}{eN_d(x)t_{eq}^2(x)/2} = 1 \tag{21}$$

Here $t_{eq}(x) = \sqrt{2t(\frac{t}{2} + \frac{\epsilon_s}{\epsilon_{ox}} t_{ox,FP})}$, which is used to model the single sided FP structure as a symmetrical FP/semiconductor structure with equivalent depletion thickness $t_{eq}(x)$ [cm], $t_{ox,FP}$ the thickness of the oxide below the FP [cm] and L_{DE} the length of the drain-extension [cm] [9].
The general strategy now to design devices that obey Equation (21), is by choosing one parameter that can be varied as a function of x and solving it in terms of the other (constant) parameters. For example, solving Equation (21) for $N_d(x)$ and $t_{ox,FP}(x)$ gives:

$$N_d(x) = \frac{\epsilon_s}{t_{eq}^2 e}\left(\frac{BV}{L_{DE}}x - V_{FP}\right) \quad \& \quad t_{ox,FP}(x) = -\frac{\epsilon_{ox} t}{2\epsilon_s} + \frac{\epsilon_{ox}\left(\frac{BV}{L_{DE}}x - V_{FP}\right)}{eN_d t} \tag{22}$$

So, the doping or oxide thickness should increase linearly in the drain-extension when all the other parameters are held constant. In a similar fashion as the equations in (22), also equations can be derived for $V_{FP}(x)$, $\epsilon_{ox}(x)$ and $t_s(x)$ and can be consulted for example in reference [9]. It must however be noted that satisfying Equation (22) only leads to optimal RESURF when Equation (20) is satisfied. The maximum doping dose is thus limited to ~10^{12} cm^{-2} [10], [14].

The FP application is an example of a single RESURF topology because the perpendicular depletion only occurs from the top. By going to multiple RESURF topologies, the perpendicular depletion occurs from multiples sides allowing for a higher doping dose than the ~10^{12} cm^{-2}. A RESURF approach that relatively easily allows for multiple RESURF topologies will be discussed next.

2.3.2 Junction assisted RESURF

Another way to generate perpendicular field gradients in the drain-extension, is the formation of junction(s) (vertical or along the width). In a n-type LDMOS device for example, this can be accomplished by implanting one or multiple p-domains in the drain-extension. In order for ideal RESURF to apply in such devices, the total charge in the p-domains should be equal to the total charge in the n-domains [1], [10]. Structures employing this junction assisted RESURF in the drain-extension are therefore often also called charge balance structures. To illustrate this concept, a simulation is performed of a simplified p-body/n-drift/n-drain structure with and without an adjacent p-domain in the drain-extension. Note that the adjacent p-domain in the drain-extension is equally sized (thickness t) and doped as the n-drift domain.

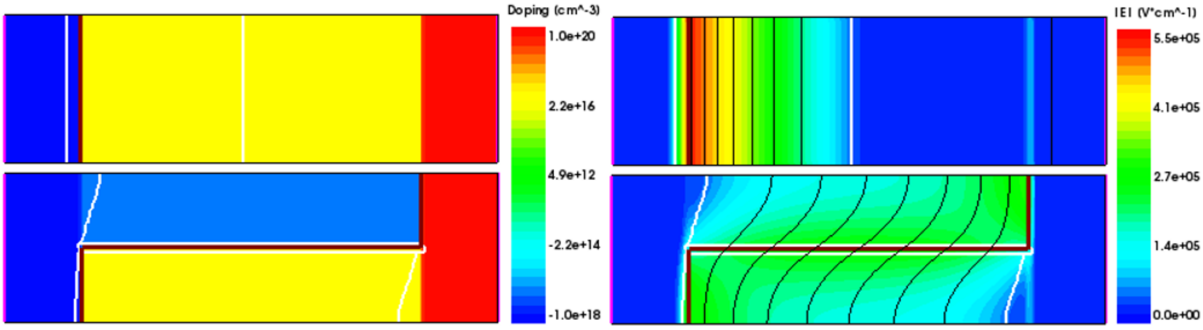

Figure 10: Simulation of a simplified p-body/n-drift/n-drain structure with (bottom) and without (top) adjacent p-domain. The body doping is 10^{18} cm^{-3}, the drift doping 10^{17} cm^{-3} and the drain doping 10^{20} cm^{-3}. The applied bias at the drain (right contact) is 10V. Left: Doping distribution. Right: Magnitude of electric field and equipotential lines. Note that the white lines indicate the depletion boundaries and the brown lines the junction boundaries.

From Figure 10, it can be seen that by adding the adjacent p-domain, the field distribution in the drain-extension becomes near uniform (deviations due to the additional depletion by the drain). This is because the n-drift region gets depleted by the adjacent p-domain resulting in a vertical field gradient. If the vertical depletion fully extends over the drain-extension thickness (i.e. $x_n = t$), all the charge in the drain-extension is used to generate the vertical field such that the horizontal field gradient must be zero. Note that this full vertical depletion must occur before the (vertical) critical field at the junction is reached, such that the doping dose in the drain-extension again must be below ~10^{12} cm^{-2} (Equation (20)) [10].

The maximum doping dose can be increased by going to architectures with multiple junctions. In the figure below, an overview is shown of the different possibilities for using junction assisted RESURF:

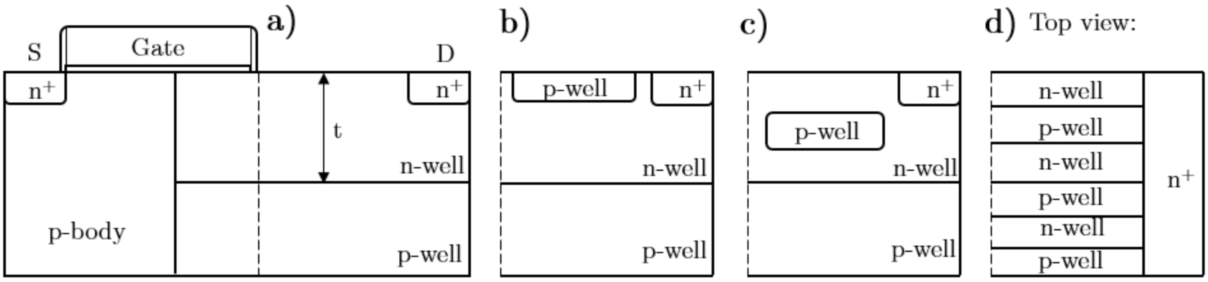

Figure 11: Schematic illustration of various possibilities for junction assisted RESURF. a) Single RESURF b) Double RESURF c) Triple RESURF d) Super junction RESURF. [1]

The configuration in Figure 11a with an implanted p-well below the n-well drain-extension, is the basic topology for most industrial LDMOS devices. Apart from the additional partial RESURF from the gate overlap, this is similar to the configuration in Figure 10. In Figure 11b, an additional p-well is implanted at the surface such that the n-well gets depleted from the top and bottom allowing for ~$2x10^{12}$ cm^{-2} (double RESURF). In Figure 11c, the additional p-well is implanted in the middle of the n-well such that the n-well gets depleted from three vertical junctions allowing for ~$3x10^{12}$ cm^{-2} (triple RESURF). The theoretical limit for this approach can be reached by the formation of a periodic structure (N periods) of alternating n- and p-domains, which can be accomplished by making use of the third dimension (Figure 11d). Such a stack of junctions is generally called a super junction and allows for ~$2Nx10^{12}$ cm^{-2}.

Although this method can be used to go to higher doping concentrations and does not require any graded parameters to obtain ideal RESURF as in the FP method, it still suffers from some drawbacks becoming more severe towards the super junction application. The most significant one, is the loss of active area through which current can flow because of the p-domains which cannot conduct the electron current in the on-state. Moreover, the additional p-domains also cause a so-called Junction Field-Effect Transistor (JFET) resistance, which is the reduction of active area with increasing bias due to the depletion of the n-drift region. Besides these theoretical limitations, there are also process limitations/difficulties regarding the implants. For example, deep and/or narrow doping domains are hard to produce. Moreover, the implants must be very carefully designed with limited process variation, since any unintentional charge imbalance can lead to large reductions in the BV. [1]

Now that an introduction to RESURF is given, along with a discussion of the two most used ways to use it, it will be shown what is the current industry standard and how this compares to the (non-)RESURF silicon device limits in the range up to 50V off-state breakdown.

2.3.3 Overview current industry standard

For power devices, it is common practice to show their performance on a $R_{ON}A$-BV graph. In Figure 12, such a graph is shown for the best performing integrated power devices in the range up to 50V extracted from literature. Most of the devices in this range typically use a single RESURF structure, for example a FP or a single vertical junction. Therefore, the lower boundary of this data is fitted to the single RESURF dependency (Equation (19)) with a correction for non-ideal RESURF by a factor ξ and for contributions to the $R_{ON}A$ other than the drift region by an offset. This "state-of-the-art fit", with ξ=1.25 and offset=0.75 $m\Omega \cdot mm^2$, is further used as a benchmark in this thesis. Moreover, to indicate the room for improvement, the theoretical limits for 1D silicon devices as well as for silicon devices with single/double RESURF topologies are shown. For calculating these theoretical limits, $d_{eff} = t = 0.3$ µm is used in Equations (16), (19) and (20).

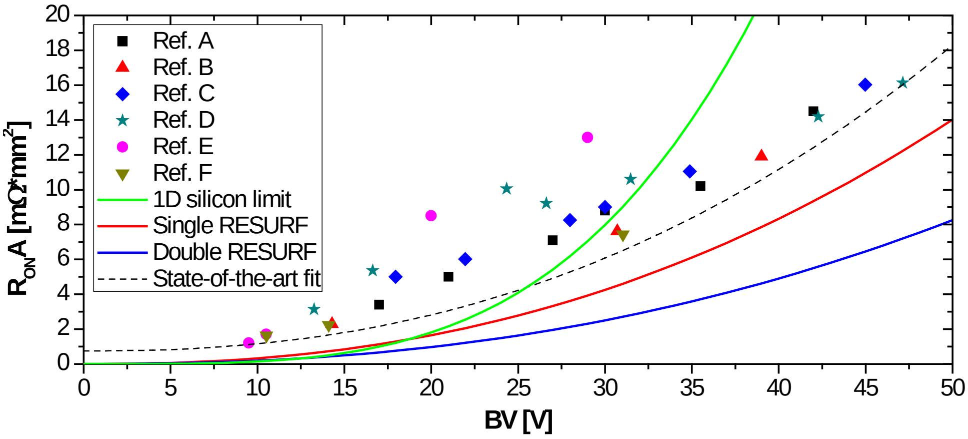

Figure 12: R_{ON}A-BV data for current state-of-the-art integrated power devices in the range up to 50V together with a "state-of-the art fit" and theoretical limits for 1D silicon devices and single/double RESURF silicon devices. The references indicated in the legend correspond to: Ref. A = HHGrace [58]*, Ref. B = TSMC* [59]*, Ref. C = MagnaChip* [60]*, Ref. D = Rohm* [61]*, Ref. E = Toshiba* [62] *and Ref. F = Renesas* [63].

From Figure 12, it can be seen that the devices known from literature only start to beat the 1D silicon limit from about ~25V. This is mainly because contributions other than the drift region, which relatively become larger for smaller BV, are not incorporated in the theoretical limits. For BV's of 10, 20, 30, 40 and 50V, the offset makes up respectively 65, 27, 12, 7 and 4% of the R_{ON}A in the state-of-the art fit. Indicating that the other contributions cannot be ignored in the lower end of this spectrum. Also, it is evident from Figure 12 that the state-of-the devices are nowhere close to the RESURF limits. Besides again the offset, this is due to non-idealities in for example the: assumed bulk mobility, geometry of junction boundaries, doping concentration/gradient and oxide thickness/gradient [10]. These non-idealities become less severe and/or can be better controlled for larger voltage applications, such that the RESURF limits typically only get approached for BV's $\gtrsim$ 100V [1], [10].

Besides trying to approach ideal RESURF conditions in the drain-extension, additional improvements can be achieved by reducing the offset. This can especially be very advantageous in the lower voltage range. Probably the most effective way this can be done, is by optimizing the channel, as this is typically the most dominant contributor to the offset.

2.4 Improving on-state resistance - channel mobility

Optimizing the channel can be done by improving L_{ch}, V_T, C_{ox} and/or μ_{ch} (Equation(7)). From these, the most preferred path would be a reduction in L_{ch} as this additionally reduces the pitch. This however, is often largely constrained by design rules (e.g. the channel should not be too short to suppress short channel effects) and existing process technologies (e.g. lithographic limits). Such constraints typically also limit optimization of the channel via the V_T and/or C_{ox}. For example, the V_T has to meet specific requirements for IC design and the C_{ox} is limited by the strict use of SiO_2 as gate oxide over other superior dielectric materials due to the existing process technology. Improving the channel mobility on the other hand, can be accomplished by the smart use/integration of existing processing technologies in an existing technology platform and is not necessarily restricted by design rules. The general expression of the mobility in a material is given by: [34]

$$\mu = \frac{e\langle\tau\rangle}{m^*} = \frac{e}{m^*}\left\langle\frac{1}{\sum_i 1/\tau_i}\right\rangle \tag{23}$$

Here $\langle\tau\rangle$ represent the mean time between successive scattering events [s], τ_i the relaxation time for a particular scattering mechanism [s] and m^* the effective conductivity carrier mass [kg].

A successful approach for lowering μ would thus both include lowering the scattering and the effective mass. It must however be noted that lowering the effective mass also lowers the carrier concentration, since the density of states (DOS) scales with the effective mass. In particular, for bulk $DOS_{3D} \propto m^{*\frac{3}{2}}$ and for a 2D system (such as the inversion layer in a MOSFET) in the quantum limit $DOS_{2D} \propto m^*$ [38]. Therefore, there will generally be an optimum effective mass [34]. Other than the effective mass, suppressing scattering is always beneficial for μ. The main scattering mechanisms in the channel of a MOSFET are phonon-, surface roughness- and Coulomb scattering (Figure 13 (left)) [34], [64]. These mechanisms are often modeled as a function of the effective transverse electric field (E_{eff} [V/cm]): [51], [65]

$$E_{eff} = \frac{1}{\epsilon_s}(\eta|Q_{inv}| + |Q_b|) \approx \begin{cases} (V_{GS} + V_T + 0.2)/6t_{oxe} & \text{for electrons} \\ -(V_{GS} + V_T - 0.25)/6t_{oxe} & \text{for holes} \end{cases} \tag{24}$$

Here $|Q_b|$ is the bulk depletion layer charge density [C/cm^2], η an empirical factor (0.5 for electrons and 0.33 for holes at room temperature [65]) and t_{oxe} an effective oxide thickness [cm]. The effective oxide thickness corrects for capacitances involved in the polysilicon and in the inversion layer. For gate oxides >10 nm it can be assumed that $t_{oxe} \approx t_{ox}$ [51].

For low E_{eff}, the mobility is limited by Coulomb scattering from ionized impurities, interface state charges and fixed oxide charges. As E_{eff} becomes higher, the increasing inversion charge starts to screen the Coulomb scatter points and phonon scattering becomes dominant. For even higher E_{eff}, the large velocity of the carriers towards the surface makes the surface roughness scattering limit the mobility. As a result, the effective mobility of carriers in the inversion layer follows a typical curve as shown in (Figure 13 (right)). The range where Coulomb scattering limits the mobility is called the low field mobility. The range where the phonons and the surface roughness limit the mobility is called the high field mobility, this is an universal curve independent of the doping. [14], [51], [65]

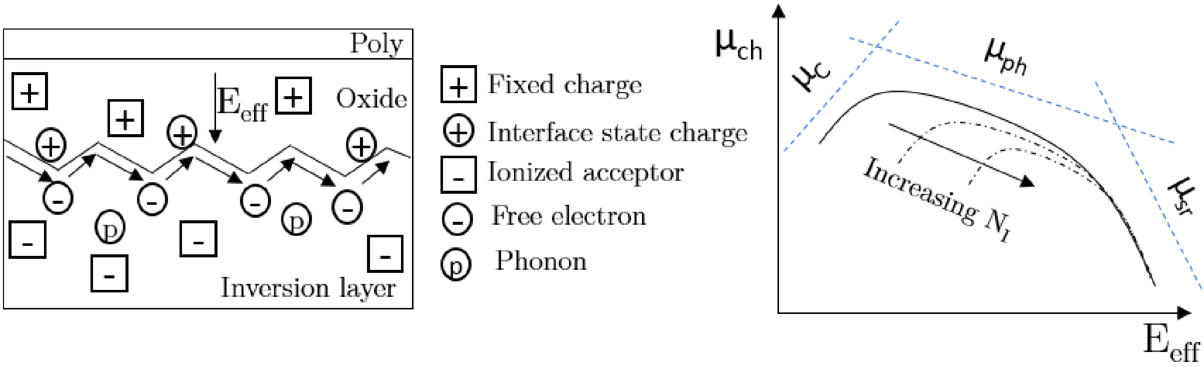

Figure 13: Left: Scattering mechanisms in the inversion layer. Right: Typical effective mobility curve of charge carriers in the inversion layer as a function of the effective transverse field. In the low field the mobility is limited by Coulomb scattering whereas in the high field it is limited by phonon and surface roughness scattering. Also, the influence of the concentration of ionized impurities (N_I) is indicated. [14], [34], [64]

One way to elevate this curve (Figure 13 (right)), would ideally be the use of superior materials over silicon, like III-V semiconductors [38]. This however, would require different process technologies, which are unlikely to be integrated in existing fabrication lines. Setting up new production lines requires large investments, both in time and money, and is therefore more a solution for the long term. Other ways to improve the mobility, that are compatible with the current silicon industry, aim to modify the band structure of silicon. This can for example be accomplished by employing strain or quantization. Before going in to depth on how this can improve the mobility, first some attention will be devoted to the band structure of silicon in the inversion layer of a MOSFET.

2.4.1 Silicon band structure in the inversion layer

Silicon is an indirect semiconductor, meaning that the top of the valence band is not aligned with the bottom of the conduction band in k-space. The lowest points in the conduction band edge of Si lie along the Δ-directions, which is approximately at 85% from the Γ-point to the X-point (zone boundary at $\frac{2\pi}{a}$<100>, where a is the lattice parameter). For bulk Si (relaxed), these valleys are six-fold degenerate (Δ_6) (Figure 14a). The valleys are highly anisotropic; the longitudinal effective mass (m_l) is approximately $0.98m_e$ while the transversal effective mass (m_t) is approximately $0.19m_e$. The constant energy surfaces of the valleys are therefore elliptic shapes. The top of the valance band is at the Γ-point (i.e. $\vec{k} = 0$), where the light hole (LH) band and the double degenerate heavy hole (HH) band cross each other. The effective masses are approximately $0.54m_e$ and $0.15m_e$ for the HH- and LH bands respectively. Below these bands, there is another band due to relativistic spin-orbit (SO) coupling. The valence band structure near Γ can then very roughly be approximated as isotropic parabolas as shown in (Figure 14b). [38]

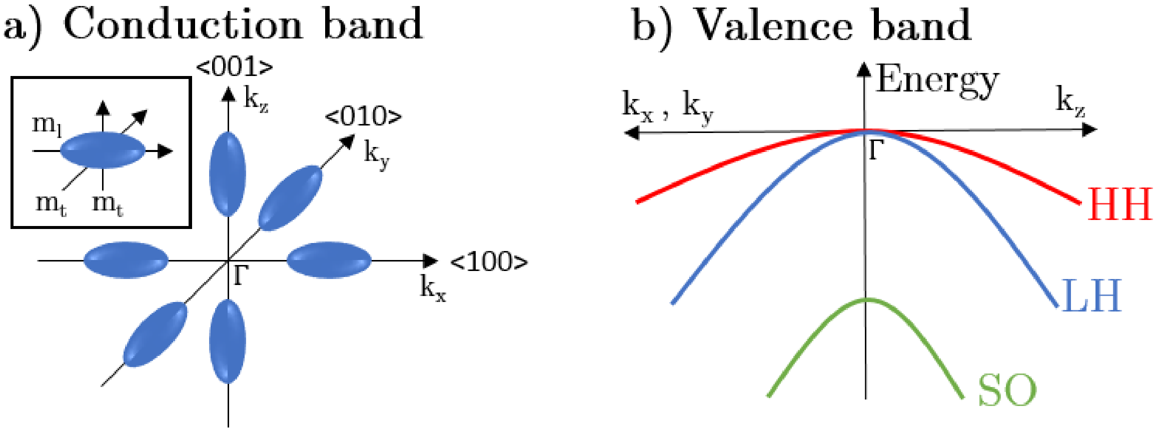

Figure 14: a) Conduction band valleys in bulk silicon. The elliptic valleys are six-fold degenerate and highly anisotropic. b) Crude isotropic parabolic approximation of the valence band structure in bulk silicon. Three different bands can be distinguished: heavy hole (HH), light hole (LH) and spin-orbit (SO). [38]

The isotropic band diagrams for both electrons and (approximately) for holes results in an isotropic conductivity tensor, which reflects the cubic symmetry of silicon. In the inversion layer of a MOSFET however, this cubic symmetry gets broken and also the degeneracy is lifted [38]. This is because the channel region of a MOSFET under strong inversion conditions cannot be regarded as being bulk silicon. The strong band bending due to the transverse electric field results in a triangular like shaped well which confines the carriers in a plane parallel to the Si-SiO$_2$ interface. In this well, the wavevector perpendicular to the interface (k_z) is quantized, leading to discrete energy levels with a 2D DOS in each level. These energy levels can be determined numerically by solving the coupled Poisson and Schrödinger equations self-consistently, although is computationally very intensive [66]. Moreover, these solutions are mainly obtained with reasonable accuracy only for n-channel MOSFET's. This is because high electric fields tend to intermingle the valence band energies, making it difficult to calculate the hole quantization [64]. Yet, an insightful approximation with predictive value can be derived analytically for the discrete energies of carriers in the inversion layer by the triangular well approach: [64], [66]

$$E_i = \left(\frac{h^2}{8m_z^*\pi^2}\right)^{\frac{1}{3}} \left(\frac{3}{2}\pi e E_{eff}\left(\frac{i+3}{4}\right)\right)^{\frac{2}{3}} \tag{25}$$

Here E_i is the energy of the i^{th} sub-band level with respect to the well [J], h the Planck constant [Js], i the energy level starting at 0 [-] and m_z^* the effective mass perpendicular to the Si-SiO$_2$ interface (z-direction) [kg].

From Equation (25), it is evident that the discrete energy levels scale inversely with the effective mass. As a result, different ladders of sub-band energies will exist for the different effective masses of the carriers, thereby lifting the degeneracy present for both electrons and holes. For electrons in particular, the Δ_6 valleys are split into two-fold degenerate valleys (Δ_2) along the k_z-direction and four-fold degenerate valleys (Δ_4) in the k_xk_y-plane of the Brillouin zone (Figure 15 (left)). The Δ_2 valleys have the heaver effective mass (m_l) perpendicular to the interface and the lighter effective mass (m_t) parallel to the interface. The Δ_4 valleys have m_t perpendicular to the interface and m_t and m_l parallel to the interface. Due to this anisotropy, the sub-band energy levels are lower and the transport mobility is higher in the Δ_2 valleys than in the Δ_4 valleys. In addition, the occupancy of the Δ_2 valleys and the Δ_4 valleys is almost the same at room temperature, due to the higher DOS of the Δ_4 valleys with respect to the Δ_2 valleys. [34], [67], [68]

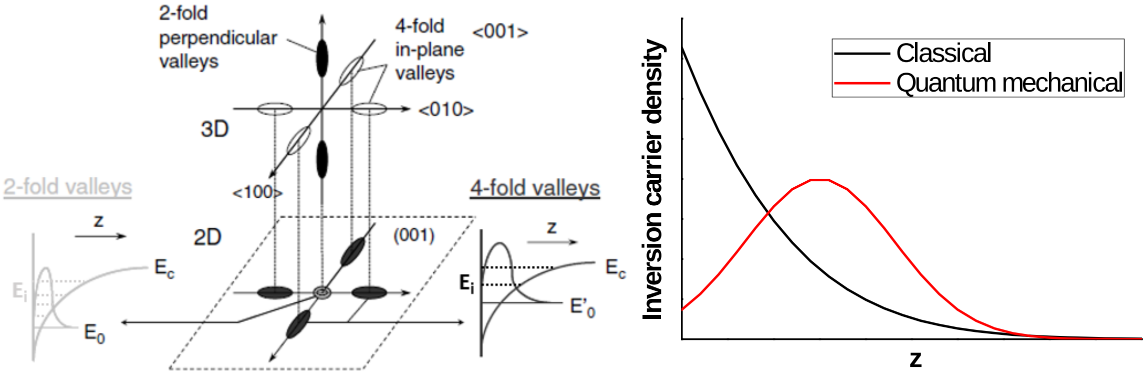

Figure 15: Left: The six-fold degenerate conduction band (Δ_6) splits into two-fold (Δ_2) and four-fold (Δ_4) degenerate valleys due to the quantization in the inversion layer. The Δ_2 valleys have the heaver effective mass (m_l) perpendicular to the interface and the lighter effective mass (m_t) parallel to the interface. The Δ_4 valleys have m_t perpendicular to the interface and m_t and m_l parallel to the interface. Due to this anisotropy, the sub-band energy levels are lower and the transport mobility is higher in the Δ_2 valleys than in the Δ_4 valleys. [67]
Right: Inversion carrier density as a function of the distance to the oxide interface from a classical and quantum mechanical treatment. Incorporating quantum effects displaces the peak towards the substrate and reduces the total inversion charge. [64]

Another important consequence of the quantum mechanical treatment of the inversion layer, can be seen in Figure 15 (right). First of all, the peak inversion charge density is effectively displaced towards the substrate due to the wavefunction boundary conditions at the oxide interface. Secondly, the total inversion charge becomes less due to the smaller DOS of a 2D system. In advanced MOSFET's with sub 90 nm channel lengths, these quantization effects have significant impact on the device characteristics (e.g. V_T and C_{ox}) [68]. At these small scales, large doping concentrations and thin gate oxides are needed to suppress short channel effects, leading to large band bending and thereby strong quantization [11], [64]. Power devices typically have larger channel lengths (>200 nm), such that less high gate induced electric fields are needed and the quantization effects are less significant (Equation (25)). Moreover, for relatively low E_{eff}, the sub-band spacing becomes small such that the 2D DOS can be approximated as a quasi 3D DOS [34], [66].

The oxygen-inserted layer (OIL) technique employs, amongst other things, the quantum mechanical properties of the inversion layer to enhance the channel mobility. This technique will be discussed next.

2.4.2 Oxygen-inserted layer technology

A potentially promising way for improving the channel mobility, is by burying an OIL close to the silicon surface [2]–[7]. This OIL consists of ordinary silicon in which partial monolayers (PM's) of oxygen are incorporated in a periodic manner. In Figure 16a (top), it can be seen how such a PM is incorporated. The buried OIL improves the channel mobility due to two mechanisms: the formation of a super-steep-retrograde-well (SSRW) doping profile and quasi-confinement of the inversion carriers. The SSRW doping profile accounts approximately 60% for the increase in the mobility and the other 40% is due to the quantum quasi-confinement effect [4]. In Figure 16b, it is shown from measurements and simulations what is the typical effect of the OIL on the electron channel mobility as a function of $|Q_{inv}|$ (equivalent to E_{eff} (Equation (24))) [3]. In addition, also a simulation is shown which solely shows the effect of the SSRW doping profile. It should be mentioned that the OIL here contains four PM's of oxygen.

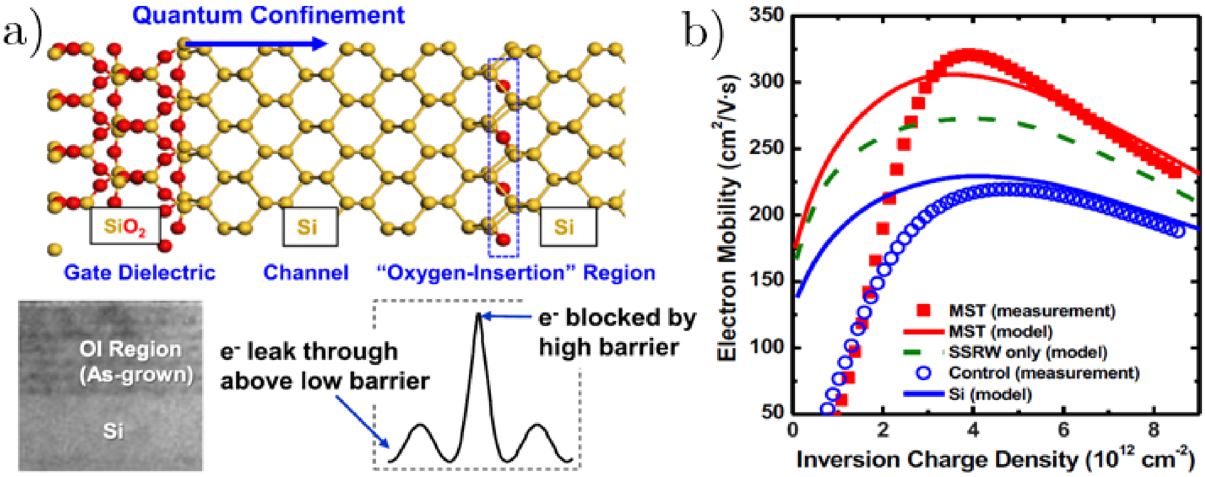

Figure 16: a) Top: Schematic view of how a PM of oxygen (blue rectangle) is incorporated below the channel region (yellow= Si atom, Red=O atom). Bottom left: TEM (Transmission Electron Microscopy) image of an OIL (superlattice of PM's of oxygen) below a small cap layer of intrinsic silicon. Bottom right: Effect of the inserted oxygen on the silicon crystal potential. b) Electron channel mobility as a function of $|Q_{inv}|$ with (MST) and without (control) OIL technology (with four PM's of oxygen) from both measurements and simulations. Also, a simulation is shown which solely shows the effect of the SSRW doping profile. The OIL results in an increase in the electron mobility mainly at the peak and high field. [4]

In order to give a quantitative impression of the mobility improvements that can be established using OIL technology, peak and high field mobility improvements are shown for both electrons and holes in Table 1 (extracted from references [4], [5]). From this table, it is evident that by using more PM's of oxygen (e.g. four instead of two), the mobility is further improved as a result of the enhanced effect on the doping and confinement [4]. How and why the OIL exactly leads to the improvements as shown in Figure 16b and Table 1, will be clarified in more detail in subsections 2.4.4 and 2.4.5. But first, it will be elaborated on how the OIL is built up, how it is integrated in the processing of devices and how it limits thermal budgets during processing.

Table 1: Indications of expected mobility improvements for OIL's with two PM's of oxygen or with four PM's of oxygen for electrons and holes at the peak and high field mobility. Four instead of two further enhances the mobility for electrons, for holes such data is not available. [4], [5]

	Electron: peak	Electron: high field	Hole: peak	Hole: high field
2 PM's of O	21%	14%	17%	8%
4 PM's of O	45%	24%	-	-

2.4.3 OIL technology - Processing

The OIL consists of a superlattice of monolayers of silicon with and without interstitially inserted oxygen (O_i) (Figure 16a (bottom left)). These O_i atoms occupy bonded off-axis interstitial sites with a mean position between nearest neighbor silicon atoms (Figure 17 (left)) and are electrically neutral [69]. The periodicity of the repeating structure and the occupation of the possible O_i sites in the PM's, provide ways to quantum engineer the brand structure. The repeating superlattice structure typically consists of seven monolayers of silicon or less, interleaved by one PM of oxygen. The preferred occupation of O_i in the PM's of oxygen is in the range from about one-eighth to one-half of the possible O_i sites. [43]

Integration of the OIL takes place after the standard shallow trench isolation (STI) implementation and implant doping processes. As a first step, the top layer of the doped silicon is etched by approximately 30 nm to obtain a clear surface required for further epitaxy. Then, using (selective) epitaxy the OIL is grown in a low temperature (<800 °C) epi process [70]. This is accomplished by intermittently exposing the silicon surface to a dilute oxygen source. The oxygen atoms are then situated on non-substitution sites between Si-Si bonds (Figure 17 (left)) which minimizes disruption of the silicon lattice and does not hinder further silicon epitaxy. As a final step, a cap layer of undoped silicon is grown on top of the OIL, where the conducting channel takes place. The thicknesses of the cap and OIL depend on the application. In this work, they are both approximately 10 nm. [2]–[7]

For standard CMOS processing, the rest of the fabrication after the OIL integration follows the standard process flow. For power devices however, caution must be paid to the applied thermal budget after the OIL integration. A too high thermal budget for a significant amount of time can result in out diffusion of the O_i atoms in the OIL, thereby losing the benefits. Diffusion is generally described by Fick's law, which contains a diffusion coefficient that captures the efficiency of the diffusion. This diffusion coefficient has a temperature dependency similar to the Arrhenius equation: [69], [71], [72]

$$D = D_0 \cdot e^{-\frac{E_a}{k_B T}} \tag{26}$$

Here D is the diffusion coefficient [cm^{-2}s^{-1}], D_0 a diffusion constant [cm^{-2}s^{-1}], $k_B T$ the thermal energy [eV] and E_a the activation energy [eV]. For the diffusion of a single isolated O_i atom in silicon, D_0 is given by 0.13 cm^{-2}s^{-1} and E_a by 2.53 eV [71], [72]. This 2.53 eV is the energy associated with crossing the energy barrier to hop from one bond-center site to one of the six other bond-center sites. In Figure 17 (right), this equation is plotted for a single isolated O_i atom in silicon (black line). Here it can be seen that the O_i diffusion is negligible before 800 °C and starts to rise rapidly afterwards.

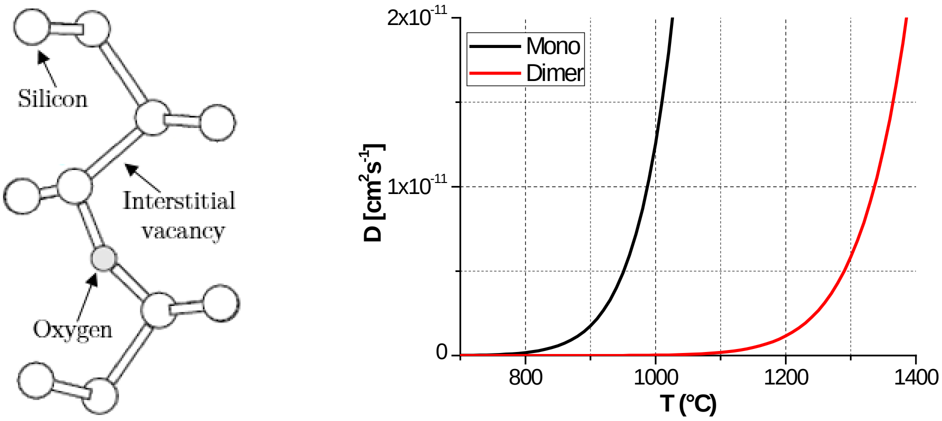

Figure 17: Left: Schematic representation of a single isolated interstitial oxygen (O_i) atom in the silicon lattice. Here also an interstitial vacancy is indicated where the oxygen atom could hop to or which could be occupied by another oxygen atom forming a dimer. [69] *Right: The diffusion coefficient for O_i atoms in silicon as a function of the temperature (black line: no interaction with other interstitials included, red line: interaction with a single other nearest neighbor O_i atom included).*

In the OIL however, the thermodynamic metastable states are stabilized due to neighbor-neighbor interaction between the multiple O_i atoms. For example, consider the situation when another O_i atom is situated at the interstitial vacancy pointed out in Figure 17 (left). Since oxygen has a large electronegativity, the O_i atoms draw negative charge from their silicon neighbors. As a result, their shared silicon atom becomes highly positively polarized leading to a quadrupole interaction between the two O_i atoms. This binding energy is predicted to be in the range of 0.7 eV to 1.7 eV [69]. Adding the lower boundary (0.7 eV) to the activation energy (2.53 eV) and again plotting the diffusion coefficient, gives the red line in Figure 17 (right). Now it can be seen that the O_i diffusion is negligible until 1100 °C and starts to rise rapidly afterwards. When multiple O_i atoms are present, such as in the OIL, the binding energy is very dependent on the relative occupation and configuration/distribution of the O_i atoms. Generally, the stabilization becomes stronger as the relative occupation increases towards one due to increased neighbor-neighbor interactions [69]. Therefore, disregarding of the exact binding energy of O_i atoms in the OIL, the out diffusion of O_i atoms can certainly be neglected below 800 °C and probably would become problematic somewhere near 1100 °C.

In the next subsection, it will be elaborated on how the OIL improves the channel mobility by lowering the channel doping.

2.4.4 OIL technology - Super-Steep-Retrograde-Well doping profile

Generally, there are three types of impurity diffusion mechanisms in silicon. The first one is vacancy diffusion, in which the impurity atom exchanges lattice positions with a vacancy. The second one is interstitial diffusion, in which an interstitial impurity atom hops to another interstitial vacancy. Atoms diffusing according to this process are generally more mobile due to the higher number of interstitial vacancies with respect to substitutional vacancies. The last one is interstitialcy diffusion, which is an interstitial mediated diffusion process. In this process, a silicon self-interstitial knocks a substitutional impurity into an interstitial vacancy, which then may knock a silicon atom into another self-interstitial vacancy (Figure 18 (left). As one could imagine, the probability for this process to occur is very dependent on the concentration of self-interstitials. In processing, there are generally two processes that generate these self-interstitials: oxidations and implants. During oxidations, approximately 1:1000 silicon atoms are unreacted which break free of the interface and diffuse interstitially in the silicon. Ion implantation generates self-interstitials by knocking silicon atoms from their lattice position, leaving a vacancy behind. Since these two processes are extensively used in device fabrication, large concentrations of self-interstitials typically exist during processing. As a result, interstitialcy diffusion may become the dominant diffusion mechanism (especially for boron and phosphorus). The number of mobile dopants then increases at the cost of substitutional dopants, resulting in enhanced diffusions of the dopants. This is also called transient enhanced diffusion (TED). [73], [74]

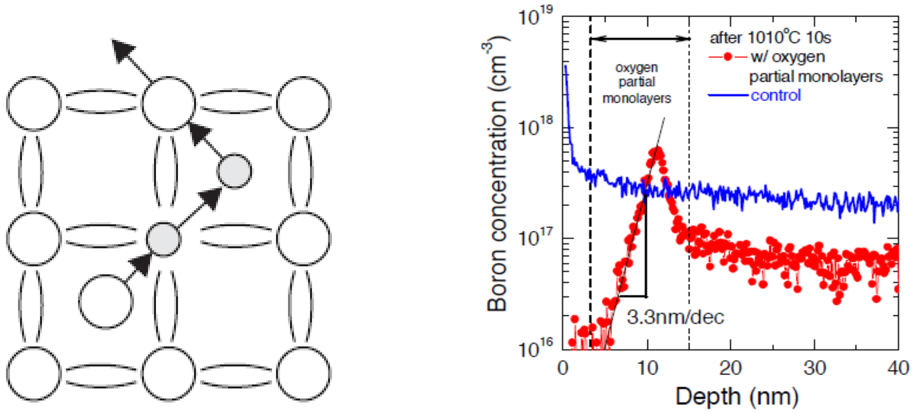

Figure 18: Left: Illustration of interstitialcy diffusion mechanism (white = silicon atom, grey = impurity atom). A silicon self-interstitial knocks a substitutional impurity into an interstitial vacancy, which then may knock a silicon atom into another self-interstitial vacancy. [74] *Right: Concentration of boron as a function of the distance to the Si-SiO$_2$ interface (obtained by SIMS (Secondary Ion Mass Spectroscopy)) through the channel region of a n-type MOSFET with (red) and without (blue) an OIL* [3]. *The OIL aids the formation of a SSRW doping profile, thereby significantly reducing the channel doping.*

Now to get back to the OIL, it is experimentally demonstrated that the PM's of oxygen block the diffusion of silicon self-interstitials and thereby effectively reduce interstitialcy diffusion beyond the depth of the OIL [70], [75], [76]. This blocking is subscribed to local charge modulation induced by the O_i atoms in the silicon lattice [3]. So, during an anneal step, doping atoms diffuse to the lower most PM where they will pile up and be impeded from further diffusion upwards to the undoped cap layer. As a result, a SSRW doping profile forms in the channel region. To illustrate this, a SIMS (Secondary Ion Mass Spectroscopy) boron doping profile through the channel region of a MOSFET with and without an OIL is shown in Figure 18 (right). It should be mentioned that the formation of the SSRW doping profile is the most effective for boron and phosphorus, since these species exhibit interstitialcy diffusion as the dominant diffusion mechanism [74], [76]. Arsenic for example, is only 40% interstitialcy driven such that the effect of the OIL on its diffusion is less than for boron and phosphorus [74], [76].

The main advantage of the SSRW doping profile is that ionized impurity scattering of carriers in the inversion layer is reduced, thereby increasing (mainly) the low field mobility. Moreover, the effect on the doping is also a way to improve short channel control by reducing DIBL due to the pile-up below the OIL (acting as a halo implant) and reducing V_T variability (>50%) caused by random dopant fluctuations. An impression of the mobility improvement due to the low channel doping, can be obtained from the curves in Figure 16b. The device with OIL (MST) shows a steeper slope in the low inversion charge density regime than the device without OIL (control), which is due to the reduced impurity scattering [4]. Moreover, from the simulations, the improvement due to the SSRW doping profile at the peak and high field are respectively 19% and 13%. For the quasi-confinement effect, the improvement at the peak and high field are respectively 12% and 10%. This effect will be discussed next. [2]–[7]

2.4.5 OIL technology - Quasi-confinement

The O_i atoms inserted in the silicon lattice act as local perturbations to the crystal potential, as illustrated in Figure 16a (bottom right). Due to this local increased potential, electrons are blocked in the vicinity of the O_i atoms but can leak through the PM's of oxygen in other regions. The channel carriers are therefore said to be quasi-confined; they experience a quantum mechanical confinement effect while the wavefunctions can still exist continuously through the material. This quantum mechanical quasi-confinement effect can be best described by thinking of the effect of the inserted oxygen on the energy band diagram below the gate oxide. Generally, increasing the amount of oxygen contained in silicon effectively widens the bandgap [43]. The actual increase in the conduction band and valence band edge in the PM's of oxygen is very dependent on the oxygen dose, nevertheless it is expected to be on the order of ~0.3 - 0.9 eV [3]–[5]. Therefore, the PM's containing the inserted oxygen can be modeled as thin layers (~1.5 nm [5]) with an increased bandgap. In Figure 19 (left), it can be seen from a simulation how two PM's of oxygen below the channel affect the lowest sub-band energies in a n-type MOSFET. The first Δ_2 sub-band decreases in energy while the second Δ_2 and first Δ_4 sub-bands increase in energy. This leads to an overall increased electron population of the Δ_2 valleys, which have the lower associated transport mass thereby decreasing the average m^* and increasing the average mobility (Equation (23)). This repopulation also decreases the gate leakage (~30-50%), as the effective mass perpendicular to the interface is larger for the Δ_2 valleys. For the p-type MOSFET, the repopulation effect is not observed [5].

Besides the effect of the PM's of oxygen on the sub-band energies, the corresponding sub-band wave functions are also affected. In Figure 19 (right), the inversion electron density distributions for the corresponding sub-bands, based on a self-consistent Poisson-Schrödinger solver, are shown for a n-type MOSFET with and without the two PM's of oxygen. [3]–[7]

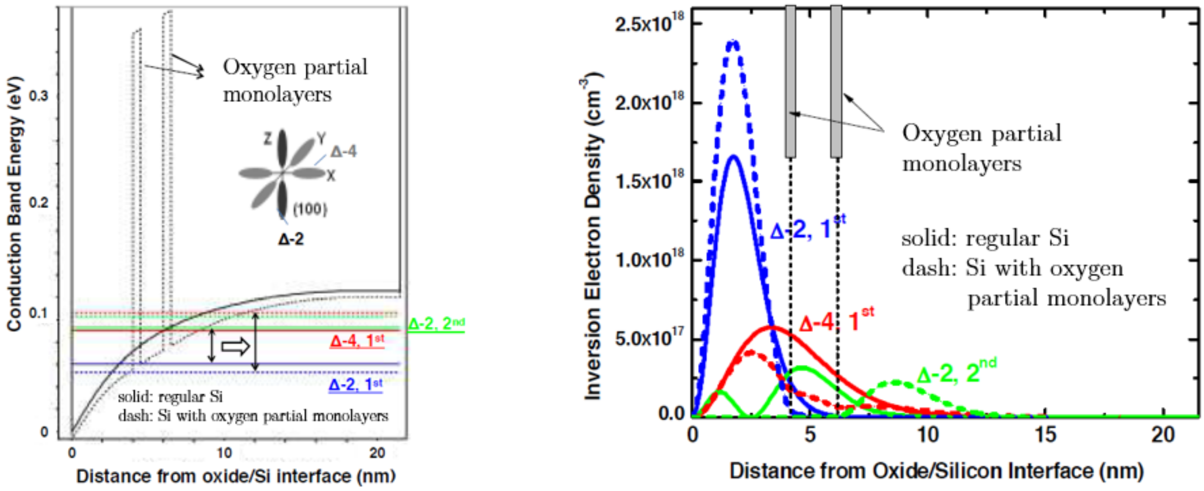

Figure 19: Left: Simulated energy band diagram and sub-band energies below the gate oxide with and without OIL (with 2 PM's of oxygen) for a n-type MOSFET. The first Δ_2 sub-band decreases in energy while the second Δ_2 and first Δ_4 sub-bands increase in energy, leading to an overall increased electron population of the Δ_2 valleys. Right: Simulated inversion electron density distributions for the first three sub-bands. A large spatial shift is observed for the second Δ_2 sub-band, thereby reducing inter sub-band scattering. [3]

From Figure 19 (right), it can be seen that besides the repopulation, the PM's of oxygen also affect the spatial separation between the carrier sub-band wavefunctions. In particular, a large shift is observed in the second Δ_2 sub-band wavefunction. For holes, a large shift is observed in the first LH-SO sub-band wavefunction [5]. This is claimed to reduce the overlap between sub-band wavefunctions [3]–[5]. Scattering between states caused by perturbations to the crystal potential (such as ionized impurities, phonons and surface roughness) can generally be described by Fermi's golden rule: [38]

$$\begin{aligned} W_{fi} &= \frac{2\pi}{\hbar}|V_{fi}|^2 N(\varepsilon_f) \\ V_{fi} &= <\psi_f|\,\hat{V}(\vec{\boldsymbol{r}},t)\,|\psi_i> = \int_{-\infty}^{\infty}\psi_f^*\hat{V}(\vec{\boldsymbol{r}},t)\psi_i\,d\vec{\boldsymbol{r}} \end{aligned} \tag{27}$$

Here W_{fi} is the transition probability per unit time, $\hbar$ the reduced Planck constant, ε_f the energy at the final state, $N(\varepsilon_f)$ the density of states in the final state, V_{fi} the matrix element, ψ_f and ψ_i the wavefunctions of the final and initial state respectively and $\hat{V}(\vec{\boldsymbol{r}},t)$ the perturbing potential.

So, from Equation (27) it is evident that reduced overlap between carrier sub-band wavefunctions, lowers the magnitude of the corresponding matrix elements and thus lowers inter sub-band scattering events. So, besides the reduction of the average m^*, the increase in mobility is also claimed to be caused by an increase in $\langle\tau\rangle$ (Equation (23)). [3]–[7]

In this section, the description of the silicon band structure (subsection 2.4.1) and the application of the OIL to increase the mobility (subsection 2.4.2 - 2.4.5) was fully devoted to the inversion channel of a MOSFET. In this work however, the OIL technique will be explored for the first time for LDMOS power devices. For these devices, the OIL has additional influences on the R_{acc} and R_{drift}. In the next subsection, it will be discussed (purely hypothetical) how the OIL could influence the R_{ON} of a LDMOS device.

2.4.6 OIL technology - LDMOS

After the current passes through the inversion channel of a LDMOS device in the on-state, it gets injected in the accumulation region. Here a part of the current flows through the accumulation layer and a part starts diffusing into the bulk. When the current is injected in the drift region, most of the current diffuses into the bulk and eventually gets collected by the drain. In the figure below, these typical current paths are indicated. Also, it is shown here how the OIL (dotted lines) would be integrated. Note that the OIL is integrated over the full device, since in practice it is integrated over the full wafer.

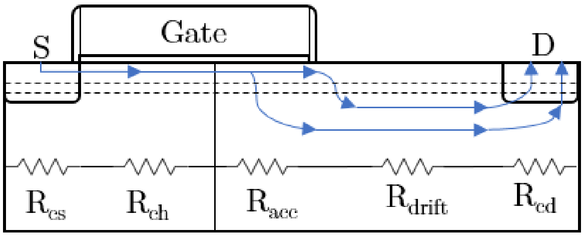

Figure 20: Typical current paths in a LDMOS device in the on-state. The dotted lines indicate how the OIL would be integrated. The layer is expected to be beneficial for the current flow through inversed/accumulated layers and its effect on the current flow into and through the bulk is hard to predict without experimental data.

The influence of the OIL on the R_{ON} is determined by its influence on the independent resistance contributions. The source and drain resistances are relatively small, such that the impact of the OIL on these can be neglected. The inversion channel is just the same as that of a standard MOSFET, such that the same improvements discussed in subsection 2.4.2 can be expected. The effect of the OIL on the R_{acc} and R_{drift} however, is not experimentally confirmed yet. The current in the drain-extension, can be subdivided in a part flowing through the accumulation layer and a part flowing through the bulk. In order to perform an educated guess, the effect of the OIL along these current paths should be considered.

In the accumulation layer, as well as everywhere else above the OIL, the doping is reduced by the self-interstitial blocking described in 2.4.4. Therefore, the increase in the accumulation layer mobility due to the doping effect is expected to be similar to that of the inversion layer mobility. For the quasi-confinement effect, the effect of the OIL is less evident. Just as like in the inversion layer, the strong electric field in the accumulation layer (for $V_{GS} > V_{FB}$) creates a triangular-like shaped well in which the energy levels are quantized [77]. A major difference with the inversion layer however, is that it is free of space charge. Therefore, not only the quantized carriers have to be considered as mobile charge carriers, but also the non-quantized carriers with a continuous energy spectrum. The wave functions of these carriers, which are travelling waves in the bulk, are distorted by the potential at the surface, thereby contributing to the self-consistent solution in the accumulation layer [77]. This makes it a very complex problem, however can be greatly reduced by only considering the $T \rightarrow 0$ limit. In this limit, it shown that two ladders of sub-band energies exist in the accumulation layer, just as in the inversion layer [77]. Therefore, it is expected that the quasi-confinement effect will have a similar effect on the mobility of the quantized carriers in the accumulation layer as in the inversion layer.

For the current diffusing into and through the bulk, the OIL could be advantageous as well as disadvantageous. The current that diffuses from the surface to the bulk (or otherwise) must pass through the OIL. How this alters the current (flow), depends on the current conducting properties in the OIL which in turn depends on how the OIL is engineered (e.g. oxygen dose and periodicity superlattice) [43]. For example, for the OIL design studied in reference [78] the mobility parallel to the plane is increased whereas it decreases perpendicular to the plane, probably leading to less diffusion into the bulk. This could lead to an increase in the resistance because the bulk mobility is typically higher than the mobility at the interface and the bulk resistance scales inversely with the effective depth (d_{eff}) (Equation (10)). A feature of the OIL that may be beneficial, is the pile-up of doping below the OIL. Current flowing through this region of increased doping may experience a lower bulk resistance (Equation (10)).

So in short, the OIL is expected to improve the current flow through inversed/accumulated charge layers and its effect on current flow into and through the bulk is hard to predict at this point. Experimental data of LDMOS devices with an OIL is therefore necessary to assess the net effect on the R_{ON}.

Chapter 3: Simulation methodology

Testing new devices on silicon generally requires a large amount of time and monetary investment to create the test structures on the wafers and to set up the measurement routines. The measurement results generally consist only of the I-V characteristics and the electrical parameters that can be derived from these. The device itself however, remains a black box with no easy or non-destructive ways to look into it. Therefore, when unexpected behavior is in encountered in the I-V characteristics, it can be beneficial to study the physics in the device using a simulation tool like Technology Computer-Aided Design (TCAD). TCAD is a simulation tool based on the finite element method, where calculations of interacting physical models are performed on a user-defined mesh. Therefore, it can be used as a computer based numerical framework to design devices and study their operation while also providing the ability to look into the device. The negative aspect of these finite element calculations however, depending on the used mesh and the complexity of the models, is that it can be very time consuming.

The benefit of specialized TCAD tools over general finite element simulation tools is the degree of user input. Within TCAD, many physical models are included which can be arbitrary disabled or enabled by the user. This feature can for example be handy to identify which physical process affects a certain behavior in the device. Following this approach, the most important physical models can be identified, which then in turn can be used to create general physical based analytical models. Setting up these types of models generally requires a large time investment, but once implemented they are usually computationally fast. As such they provide strong capabilities to optimize the device architecture and to predict behavior across wide design and operating ranges.

The TCAD tool used in this work is Sentaurus™ from Synopsis. In this tool, the simulation is divided into two parts: the process simulation and the device simulation. In this chapter, it will be discussed how this is built up and how it is used in this work.

3.1 Process simulation

The process simulation is the simulation of the fabrication process from the wafer substrate to the point where the electrical contacts are deposited. This is done using the Synopsys Sentaurus™ Process tool, which is an advanced 1D, 2D and 3D process simulator suitable for silicon and non-silicon semiconductor devices. All standard process steps can be simulated using this tool, such as: diffusion/annealing, ion implantation, oxidation, etching, deposition and sillicidation. Moreover, to stay as close as possible to the actual fabrication process an advanced set of built-in calibrated parameters is available. The input of Sentaurus Process is a command file which consist of a sequence of commands that corresponds to the individual process steps [79]. The devices simulated in this work are part of a standard technology platform of NXP for which TCAD input files were already available. A global overview of the process and simulation flow will be presented here.

3.1.1 Process flow

In the figure below, a simplified overview of the process flow of NXP's technology platform is shown. Note that in this overview all kinds of additional steps, such as annealing, photoresist deposition/removal and etching, are not shown.

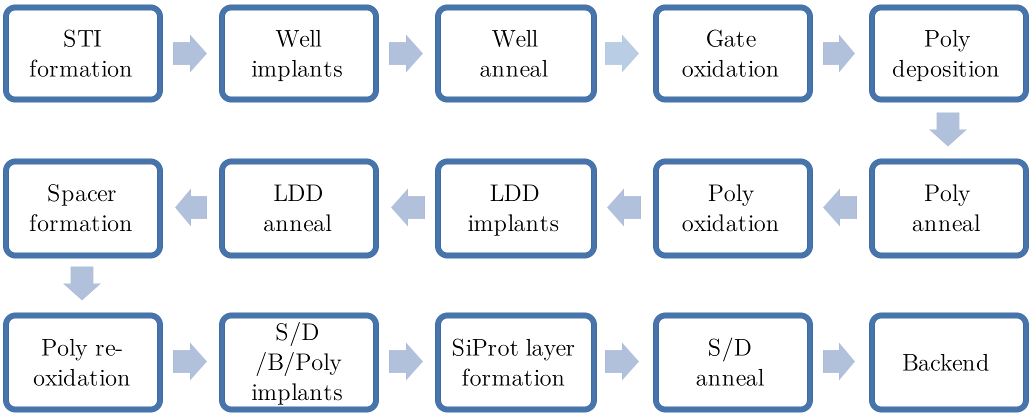

Figure 21: Simplified overview of the most important steps in the process flow of NXP's technology platform.

In this work, the wafers with the CFP-related devices exactly follow this process flow. For the wafers with the OIL devices, small adaptions are made in the implants and thermal budgets to make the oxygen insertion technique compatible with NXP's technology platform. In the next subsection, the (default) process flow will be elaborated on by means of an example of a process simulation.

3.1.2 Example process simulation

In this subsection, it will roughly be shown how the so-called Free-Analog (FA) device from NXP's technology platform is simulated using Sentaurus 2D process simulation. This device is referred to as a 12V device, indicating that it can handle up to 12V at the drain at full gate drive (5V for this device). Moreover, it is a typical example of a LDMOS device and stands at the basis of the CFP devices that are explored in this work. In the figure below, cross-sections are shown from the most important steps along the process flow simulation.

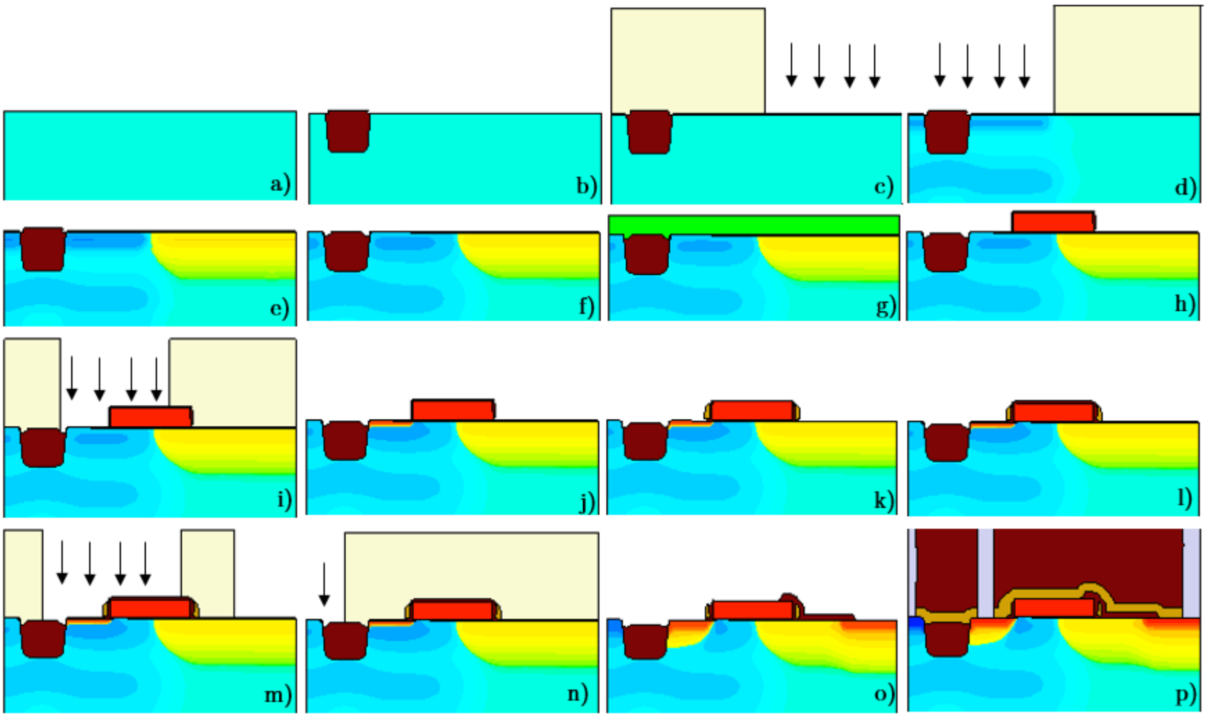

Figure 22: Process flow simulation of the FA device using Sentaurus 2D process simulation. Here the color in the silicon indicates the doping concentration (red: n-type, blue: p-type). a) Starting substrate. b) STI formation. c) Drift implants. d) Channel implants. e) Well anneal. f) Gate oxidations. g) Poly deposition and poly anneal. h) Poly etch, pre-dope and oxidation. i) LDD implants. j) LDD anneal. k) Spacer formation. l) Poly re-oxidation. m) S/D/Poly implants. n) Body implants. o) Oxidation SiProt oxide layer. p) Backend construction.

The first step in the process simulation is defining the substrate. The substrate consists of heavily boron-doped silicon (~10^{19} cm^{-3}) and a 4 μm epilayer of lower boron-doped silicon (~10^{15} cm^{-3}). Part of this epilayer is shown in Figure 22a. On this simulation domain, an initial triangular mesh is defined as well as mesh goal settings related to doping gradients, interfaces, materials and certain regions to ensure proper adaptive meshing throughout the process flow simulation. The next step is the formation of STI's (Shallow Trench Isolation). A STI is a trench etched (~350 nm) in silicon which is subsequently filled with SiO_2 (Figure 22b). These are (originally) intended to electrically isolate regions or devices. Thereafter, the well implants for the drain-extension (Figure 22c) and the channel region (Figure 22d) are performed. This is done by accelerating ionized donors/acceptors with a certain dose, energy and angle (which is defined for each implant) into the silicon while being masked by a photoresist layer.

After the implants, the dopants are thermally activated and diffused by a well anneal in excess of 1000 °C (Figure 22e). Subsequently, the gate oxidation is performed which is the growth of a SiO_2 oxide layer on top of the substrate by thermal oxidation (Figure 22f). Thermal oxidation is basically an anneal step under the influence of O_2 flow to the substrate. Oxygen then diffuses into the substrate where it reacts with silicon and forms SiO_2. This step is performed in three cycles: GO3, GO2 and GO1 oxidation, in which each cycle grows a certain oxide thickness by using different oxidation temperatures and times. Different gate oxides thicknesses are obtained by etching after the GO3 and GO2 oxidation cycles. The thickest gate oxide is denoted by GO3, intermediate by GO2 and thinnest by GO1. These have gate voltage ratings of respectively 5V, 3.3V and 1.8V. After the gate oxidation, the polysilicon for the gate is deposited and a high temperature anneal is performed (Figure 22g). This poly anneal is the highest temperature step after the well anneal and is there to mimic the thermal budget of another processes flow. In addition, it also beneficial for the quality of the gate oxide and the STI. After the polysilicon is pre-doped and etched into pattern, it is thermally oxidized by a poly oxidation to form a screen oxide for subsequent implants (Figure 22h).

Then, the lightly doped drain (LDD) implants are performed (Figure 22i). Besides the lowly doped implants near the source (and drain for symmetric devices), these contain also halo implants to prevent punch-through. These implants are activated by the subsequent rapid thermal LDD anneal (Figure 22j). Next, oxide and nitride are grown left and right of the gate and etched into spacers to isolate the gate from the source and drain (Figure 22k). After the spacer formation, the (poly) silicon is again oxidized to form a screening oxide for the subsequent implants (Figure 22l).
The last implants that are performed, are the implants for contacting the source, drain, gate and body nodes. The source, drain and gate implants are performed using the same heavy implants (Figure 22m). The body implants are performed separately since connecting the body requires opposite type of doping (Figure 22n). Thereafter, a silicide protection layer of oxide (SiProt) is deposited on top of the drift region (if present) to protect it from silicidation of the contacts. Silicidation is the process in which silicon reacts with a transition metal forming a metal silicide, which is used to form low-ohmic contacts to the silicon. The actual silicidation is not simulated as it is time consuming, prone to meshing problems and not necessary to obtain accurate solutions. In addition, the SiProt step also includes a rapid thermal S/D anneal to activate the S/D/B/Poly implants (Figure 22o). Lastly, the backend structure is constructed, which is eventually needed to control the device. Amongst others, this includes: formation of nitride capping layer, tungsten contact deposition and oxide filling (Figure 22p).

After the process simulation is done, two more things remain to be done prior to electrical simulations. First, it must be defined where the S/D/B/G contacts are. In this work, the source and drain are contacted at their tungsten contacts and the gate at the top of the polysilicon. The body can be contacted at the separate tungsten body contact for a separate source-body construction (as in Figure 22p) or is contacted together with the source for a merged source-body construction. In simulations however, the body contact is often defined directly at the silicon substrate such that the process simulation of the body contact can be omitted thereby decreasing simulation time and improving convergence for the device simulation. Secondly, the structure should be re-meshed for device simulation. This is on one hand to minimize simulation time and on the other hand to obtain physical accurate solutions. For example, the number of grid points in tungsten, oxide, nitride and deep into the silicon substrate can be reduced since these regions don't directly alter the electrical performance. The number of grid points in critical regions however, such as at junctions and at the Si-SiO$_2$ interface, should be increased to properly capture the fine details in the simulation. After the contact definitions and re-meshing, the device simulation can be performed.

3.2 Device simulation

After the device is simulated using the Sentaurus process tool, the electrical behavior can be numerically simulated using the Synopsys Sentaurus™ Device tool. Advanced physical models and robust numeric methods are incorporated to be able to simulate a wide range of semiconductor devices. Terminal voltages, currents and charges are calculated based on a set of physical device equations that describe the carrier distribution and conduction mechanisms. The set of physical device equations that are used depend on the models that are enabled or disabled by the user. In the next subsection, it will be discussed which physical models are used in this work. [47]

3.2.1 Physical models

The fundamental set of equations that must be solved to describe the electrical behavior of a device are the Poisson equation and the continuity equations: [47]

$$\epsilon_s \nabla \bullet \nabla V = -e(p - n + N_d - N_a) - \rho_{traps} \tag{28}$$

$$\nabla \bullet \overrightarrow{J_n} = e(R_n - G_n) + e\frac{\partial n}{\partial t} \quad \& \quad \nabla \bullet \overrightarrow{J_p} = e(R_p - G_p) + e\frac{\partial p}{\partial t} \tag{29}$$

Here p and n are the hole and electron concentrations respectively [cm^{-3}], N_d and N_a the ionized donor and acceptor concentrations respectively [cm^{-3}], ρ_{traps} the charge density contributed by traps and fixed charges [C/m^3], $\overrightarrow{J_n}$ and $\overrightarrow{J_p}$ the electron and hole current densities respectively [A/cm^2], R_n and R_p the electron and hole net recombination rates respectively [cm^{-3}s^{-1}] and G_n and G_p the electron and hole net generation rates respectively [cm^{-3}s^{-1}].

The current density, combined with Equation (28) and (29), can be solved self-consistently according to several models in Sentaurus device. The simplest and default model, is the drift-diffusion model. This model assumes the current density to be a function of the spatial variation in: the carrier density, the conduction/valence band edge and the effective mass. A more accurate model, which additionally solves the lattice temperature (heat flow) equation, is the thermodynamic model. In this model, an extra driving term proportional to the gradient of the lattice temperature is included in the current density. Here it is assumed that the charge carriers are in thermal equilibrium with the lattice. The hydrodynamic model drops this assumption and also solves the carrier temperature equations. In this model, an extra driving term is added to the current density proportional to the gradient of the carrier temperature. [47]

For a fairly simple device simulation, the set of differential equations (dependent on the selected transport model) are solved self-consistently and describe the electrical behavior in most of the bulk regions of the device. However, to properly simulate the device also at critical points, such as material interfaces and junctions, additional models must be adopted to give reliable results. The models used in this work are shown in the following overview:

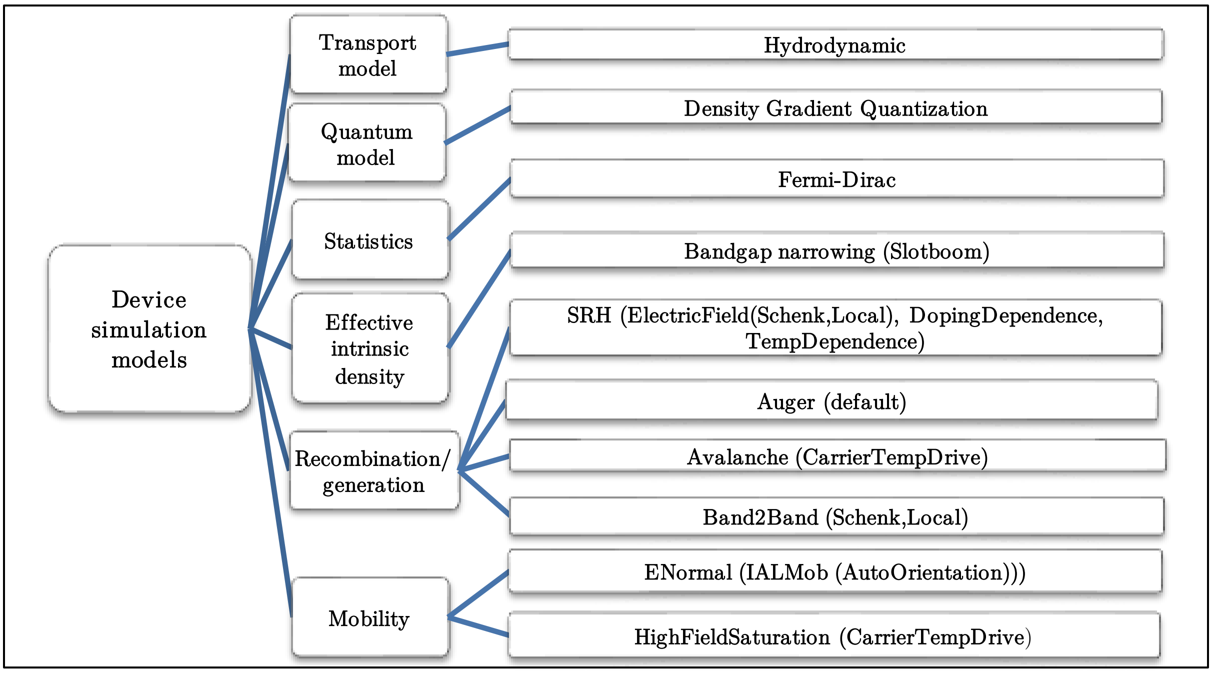
Device simulation models
Transport model
Hydrodynamic
Quantum model
Density Gradient Quantization
Statistics
Fermi-Dirac
Effective intrinsic density
Bandgap narrowing (Slotboom)
Recombination/ generation
SRH (ElectricField(Schenk,Local), DopingDependence, TempDependence)
Auger (default)
Avalanche (CarrierTempDrive)
Band2Band (Schenk,Local)
Mobility
ENormal (IALMob (AutoOrientation)))
HighFieldSaturation (CarrierTempDrive)

Figure 23: Overview of the physical models used in this work for electrical device simulations.

For the transport model, the most complete hydrodynamic model is used. Although this model is computationally very intensive, it turned out to be necessary to avoid the onset of premature breakdown in off-state breakdown simulations. For the other transport models, the driving force used for the avalanche and high-field-saturation models is the gradient of the quasi-Fermi potential. Since this assumes instant equilibrium of the carriers with the local field, this can lead to overestimation of the carrier velocity and thus premature breakdown. In the hydrodynamic model however, the effective field used as the driving force is derived from the carrier energy itself. To still win some computation time, by default self-heating is switched off by fixing the lattice temperature at 300K. For low current simulations, which are of main interest in this work (e.g. for BV and R_{ON}) , this is not relevant because self-heating is negligible. [47]

To account for quantum mechanical quantization effects on the carrier density distribution next to the Si-SiO_2 interface, the Density Gradient (DG) model is used. Besides solving the Schrödinger equation, this model gives the most physically accurate solutions and is recommended by Sentaurus for silicon. The DG model computes a so-called quantum potential which is used as a correction to the quasi-Fermi levels. The quantum potential is a function of the carrier densities and their gradients, such that it requires the self-consistent solutions with the carrier transport equations. The main reason a quantization model is used in this work, is to get the V_T similar to measurements. [47]

For the statistics, the physically correct Fermi-Dirac distribution is used. The default statistical model used in Sentaurus device is the Maxwell-Boltzmann distribution. This distribution however, is not accurate for high carrier densities ($>10^{19}$ cm^{-3}) which are typical for the inversion layer in MOSFET's [47].

If semiconductors are heavily doped ($>10^{18}$ cm^{-3}) [14], the bandgap becomes effectively more narrow. This reduction in bandgap energy at highly doped places in turn affects the intrinsic carrier density. To account for this phenomenon, the Slotboom model is used [47].

The Schokley-Read-Hall (SRH) model for recombination/generation enables additional recombination/generation through deep defect levels within the energy gap. The lifetimes of electrons and holes are modeled as functions of the doping, electric field and temperature. The doping dependence option modifies the SRH lifetimes according to the doping using the Scharfetter relation. The temperature dependence option modifies the SRH lifetimes according to the temperature using an empirical power law model. The electric field option reduces the SRH lifetimes in regions of strong electric fields (~ >30 V/µm) using the Schenk-Trap-Assisted-Tunneling Density Correction model. [47]

The Auger model activates the Auger recombination. In this type of recombination, two particles recombine and the released energy is transferred to a third particle instead of coupled to the lattice. Since this is a three particle mechanism, it becomes more important at high carrier concentrations. [47]

The Avalanche model accounts for the electron-hole pair production due to avalanche generation, as was explained in subsection 2.1.3. This is especially important for breakdown simulations. The driving force of the carriers is derived from the carrier temperature. [47]

The Band2Band model for recombination/generation accounts for band-to-band tunneling, which for example is important to simulate GIDL (subsection 2.1.3). Here the Schenk model is used with local density correction, which accounts for the non-uniform (spatial) generation of electrons and holes at the ends of the tunneling path [47].

The Inversion And Accumulation Mobility model (IALMob) is used for the mobility, which is typically used for power devices. This model has contributions from Coulomb impurity scattering (2D and 3D), phonon scattering and surface roughness scattering. As an option, Auto-Orientation is specified which switches between different parameter sets based on the crystal orientation of the nearest interface. [47]

The HighFieldSaturation model accounts for velocity saturation under the influence of high electric fields. The electric field dependence of the mobility is calculated using the extended Canali model in which the driving force is derived from the carrier temperature. [47]

Now that the choice of physical models is explained, it will be discussed how and what type of electrical simulations are performed in the device simulations.

3.2.2 Electrical simulations

Three types of standard characteristics are simulated using Sentaurus device, namely: $I_D - V_{GS}$, $I_D - V_{DS}$ and off-state breakdown. In this subsection, it will be briefly discussed for each of these simulated measurements how they are executed, what their curves typically look like and what characteristics can be derived from it.

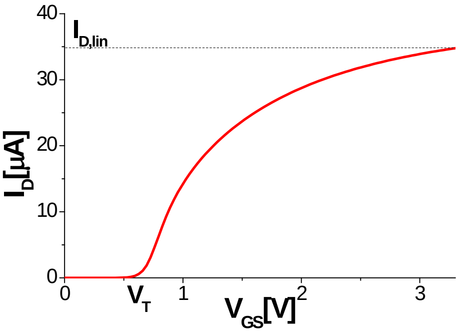

Figure 24: Example of simulated $I_D - V_{GS}$ curve.

The $I_{DS} - V_{GS}$ characteristic is simulated by ramping the drain voltage quasi-stationary to 0.1V and subsequently ramping the gate voltage to the operating voltage (1.8V for GO1, 3.3V for GO2 and 5V for GO3). The drain voltage of 0.1V is chosen such that the device operates in the linear regime. A typical simulated curve can be seen in Figure 24. The two most important characteristics that can be obtained from this curve are the R_{ON} and V_T. The R_{ON} is simply determined by dividing the drain voltage (0.1V) through the drain current at the gate operation voltage ($I_{D,lin}$). For the V_T extraction, a wide variety of methods are used in the industry [80]. In this report, the Extrapolation in the Linear Region - method is used, which extracts the so-called $V_{T,gm}$. This method consists of finding the V_{GS} - axis intercept of the linear extrapolation of the $I_D - V_{GS}$ curve at the maximum transconductance. From Equation (4), V_T can then be determined by subtracting $V_{DS}/2$ from this intercept (see also Figure 4 (left)).

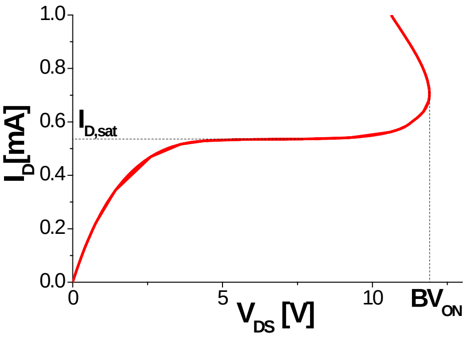

Figure 25: Example of simulated $I_D - V_{DS}$ curve.

The $I_{DS} - V_{DS}$ characteristic is simulated by first ramping the gate voltage quasi-stationary to the operating voltage and subsequently the drain voltage is ramped quasi-stationary to a voltage larger than the saturation voltage (such that on-state breakdown is visible). These types of simulations (i.e. where breakdown occurs) cannot always be simulated by simply stepping the drain voltage. Therefore, to simulate the full curve, drain current stepping is initiated after drain voltage stepping failed to converge. If also current stepping fails to converge, the so-called continuation method is used. This method is based on a dynamic load-line (connected to the electrode at which the curve is traced) which adapts the boundary conditions along the traced curve to ensure convergence [47]. A typical simulated curve using this routine is shown in Figure 25. The two most important characteristics that can be obtained from this curve are the saturation current $I_{D,sat}$ and the on-state breakdown voltage BV_{ON}.

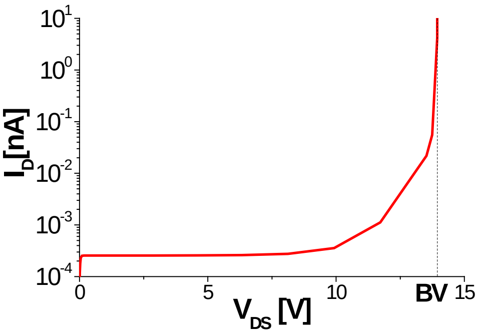

Figure 26: Example of simulated off-state breakdown curve.

The off-state breakdown is simulated in the same way as the $I_{DS} - V_{DS}$ curve, only now with zero gate voltage applied. An example of a simulated off-state breakdown curve is shown in Figure 26. This simulation is generally the hardest to perform due to the low avalanche currents. Sub-optimal meshing and/or numeric controls lead readily fast to poor convergence in the lower current regime. This in turn can lead to very large computation time or non-converging solutions. In addition, the extracted curve is also quite sensitive to the numerical settings and the meshing, such that parameters extracted from this curve should mainly be interpreted qualitatively. The most important parameter that can be extracted from this curve is the off-state breakdown voltage (BV), which is defined as the voltage where the current start to rise indefinitely (unless stated otherwise).

Chapter 4: Experimental methodology

The simulation methodology presented in the previous chapter contains many parameters and settings which can and need to be tuned to achieve accurate results for a specific process and device. Consequently, without extensive calibration and dedicated models, simulations only provide the device's characteristics up to an accuracy of about 10%. Moreover, the simulations in this work do not account for changes in the device behavior due to for example charge trapping, which may be crucial for the device reliability. Therefore, TCAD simulations can never be a complete replacement for experimental results. Simulations should only be used to obtain a first order approximation of the characteristics or to qualitatively understand what is happening inside the device. To achieve industrialization of a new device, it should always be tested on silicon, typically in multiple test phases. In the first phase, a split is set up on the wafer(s), where the processing or the properties of the devices vary over the wafer(s). Splits can for example be set up in the dimensions of the devices, in the doping implant doses and/or in the applied thermal budgets. From this first phase, it can then be mapped how the device's characteristics depend on the various process/device parameters. This knowledge can subsequently be used to design devices that need to meet certain requirements. In further test phases, these devices can optionally be further analyzed/optimized.

In this chapter, it will first be discussed what experimental setup is used to perform electrical measurements. Thereafter, it will be elaborated on which measurement routines are used to perform the basic electrical measurements, reliability measurements and capacitance measurements.

4.1 Experimental setup

In order to perform measurements on small transistors situated on silicon wafers, generally two parts of equipment are needed. First of all, a probe station, which armed with strong magnifying cameras and thin needles (probes) can be used to make electrical contact with the device under investigation (DUT). Secondly, a device analyzer, which can be used to control and to measure the voltages/currents at the probe sites. In this section, it will be discussed how these pieces of equipment are used in this work.

4.1.1 Probe station

The probe station that is used in this work is the CM300 from CascadeMicrotech® and is shown in Figure 27. The main components are the wafer chuck, eight probes (i.e. needles), two cameras for a top and side view of the wafer and two screens to visualize the wafer/probes and the location on the wafer map in real-time. In addition, the microchamber containing the wafer, is connected to a thermal control system (ERS-AC3), which keeps the temperature stable at the desired temperature (default: 300 K) during measurements. Detailed specifications can be found in reference [81].

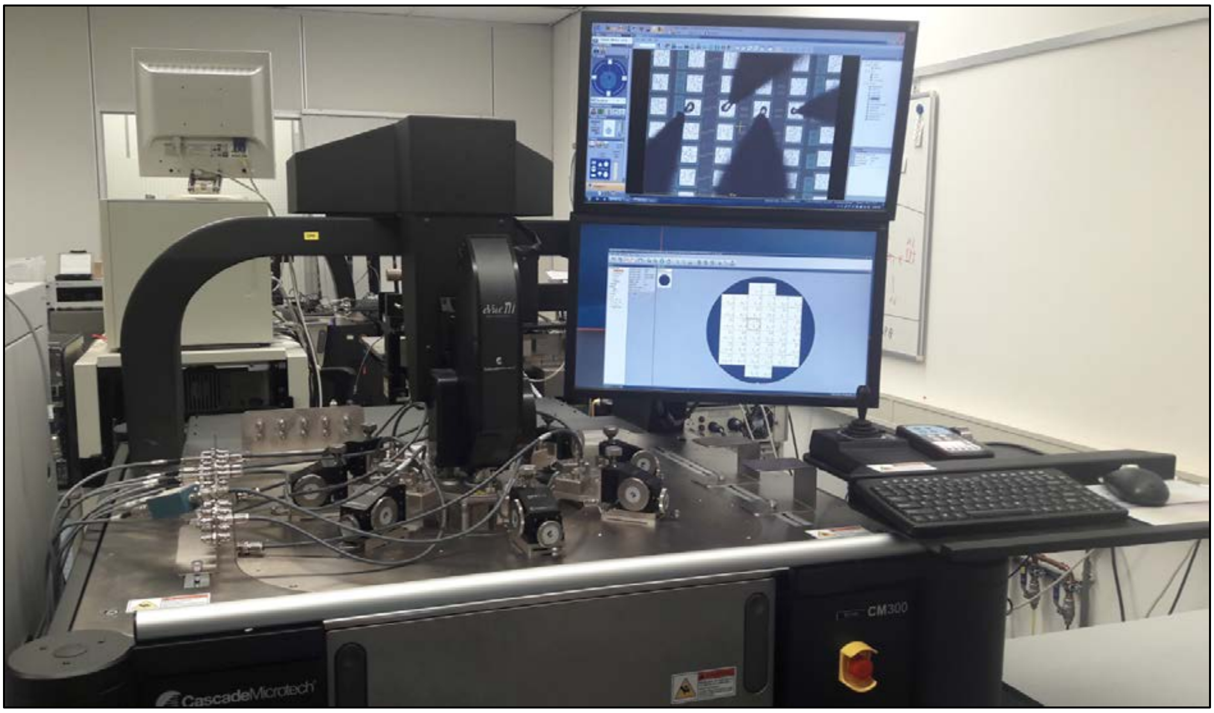

Figure 27: Probe station (CM300 from CascadeMicrotech) used in this work to perform electrical measurements.

The first step in the measurements procedure, is placing the desired wafer on the wafer chuck. The wafer is clamped to the chuck by manually activating a vacuum below the wafer. When the wafer is loaded, the Velox™ probe station control software shows a magnified part of the wafer on the top screen and the location (indicated by a dot) on the wafer map on the bottom screen. This software fully controls the cameras and the x,y- and z-stage of the wafer chuck. In order to prevent any unwanted contact between the probes and the wafer, the probe horizon can be set using the side camera. Furthermore, to align the camera movement with the wafer movement, a calibration procedure can be executed. This option is enabled by the standard calibration patterns that are etched in each die (periodic repeating structure on wafer). After the calibration procedure is successfully executed, switching between different die's/devices can then simply be performed by a built-in index stepper. Contacting the bond pads of the device with the desired probes (in contact mode) can be done using the probe manipulators. These DPP210 manipulators all have three rotary handles, to move the needles along the x-, y- and z-directions. When the needles are in contact with the bond pads, the only thing that remains is switching off the lights in the microchamber. Controlling the electrical measurement is done with the device analyzer and is discussed next.

4.1.2 Device analyzer and software

The device analyzer used in this work is the semiconductor device analyzer B1500A from Keysight technologies® and is shown in Figure 28 (left). This analyzer integrates multiple measurement resources in a single box to perform current-voltage and capacitance measurements accurately. The probes of the probe station are connected to the device analyzer via SMU's (Source-Measure-Unit) or via CMU's (Capacitance-Measure-Unit). The SMU is a measurement component that combines the capabilities of a voltage/current source and a voltage/current sensor into a single component. Likewise, for capacitance measurements a CMU can be used for both source and sensor. Due to this close integration of the source and measurement circuitry, superior measurement accuracy can be obtained with respect to using various independent instruments. The Medium-Power SMU's used in this work are capable of performing electric measurement in the range from 10 fA - 0.1 A / 0.5 µV - 100 V with a minimum source resolution of 50 fA / 25 µV. The Multi-Frequency CMU used in this work, can sent and monitor AC signals from 1kHz to 5 MHz with an amplitude range from 10 mV$_{rms}$ to 250 mV$_{rms}$. More detailed specifications can be found in reference [82].

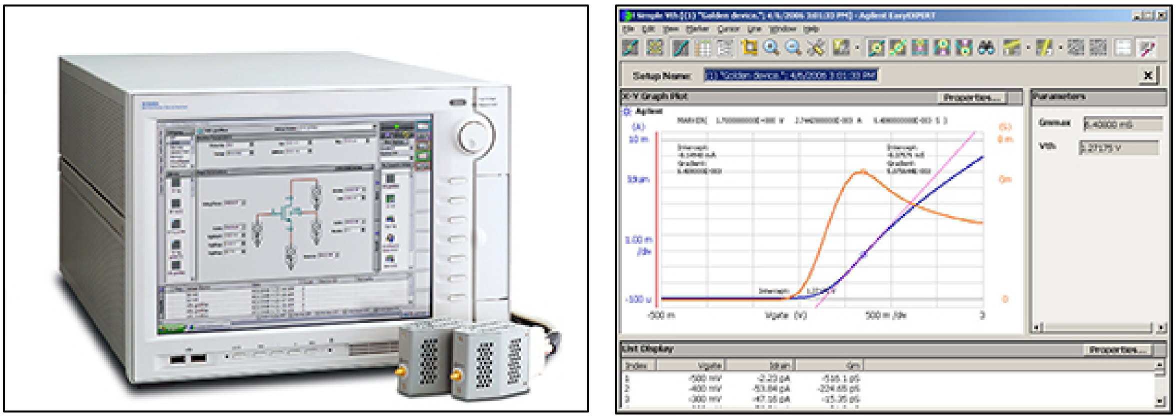

Figure 28: Left: The Keysight B1500A semiconductor device analyzer. Right: Example of $I_D - V_{GS}$ curve in EasyEXPERT and automatic $V_{T,gm}$ extraction. [82]

Controlling, extracting and analyzing the measurements can be done either with an external computer in combination with a compatible software or with the device analyzer itself. In this work, mostly the device analyzer itself is used for electrical measurements. The software that is used to set up measurements and perform subsequent analysis, is called EasyEXPERT from Keysight Technologies. Setting up a measurement using this software, starts with defining the contact names (Source/Drain/Gate/Body) to the correct SMU's. After that, it is defined which contacts are held at a constant voltage and which contact voltage is ramped in a desired number of steps. Optionally, also some built-in extraction methods can be activated. For example, in a $I_D - V_{GS}$ measurement, the $V_{T,gm}$ can be extracted automatically as is shown in Figure 28 (right). During a measurement, the software interactively displays the measured curve as well as the extracted data and therefore provides strong insight during measurement. After the measurement is done, the data can easily be extracted as a *.csv file. [83]

4.2 Measurement routines

In this work, three types of measurements are performed. First of all, the basic electrical measurements which are typically used for characterization. Secondly, reliability measurements which are used as an indication for the lifetime of a device. Lastly, capacitance measurements which can be used to say something about switching losses or to derive other important parameters such as t_{ox} and $\mu_{ch}(E_{eff})$.

4.2.1 Basic electrical measurements

The basic electrical measurements are the $I_D - V_{GS}$, $I_D - V_{DS}$ and off-state breakdown measurements. The (typical) settings that are used for these measurements are shown in the table below:

Table 2: Typical settings used for the basic electrical characterization measurements. (S=Source, B=Body, D=Drain, G=Gate)

Measurement	Gate Oxide	Constant SMU's	Variable SMU	Ramp up to
$I_D - V_{GS}$	1	S=B=0V D=0.1V	G	1.8V
	2	S=B=0V D=0.1V	G	3.3V
	3	S=B=0V D=0.1V	G	5V
$I_D - V_{DS}$	1	S=B=0V G=1.8V	D	Voltage rating
	2	S=B=0V G=3.3V	D	Voltage rating
	3	S=B=0V G=5V	D	Voltage rating
BV	1, 2 & 3	S=G=B=0V	D	2x Voltage rating

From the $I_D - V_{GS}$ measurement, the $I_{D,lin}$, V_T (i.e. $V_{T,gm}$) and R_{ON} can be extracted. From the $I_D - V_{DS}$ measurement, the $I_{D,sat}$ and $\mathrm{BV_{ON}}$ can be extracted. From the off-state breakdown measurement, the BV can be extracted.

4.2.2 Reliability measurements

A measure for reliability is the degradation in $\mathrm{R_{ON}}$ and $V_{T,gm}$ over time when the device is under worst-case stress [8]. Worst-case stress is defined as the condition with the maximum voltage rating at the drain and the gate voltage such that the impact ionization generated current is maximal. This worst-case stress condition can be found by ramping the gate voltage with the maximum voltage rating at the drain and measuring the substrate current (measure for impact ionization). The gate voltage at which the substrate current (and thus the impact ionization) is maximal, then gives the worst-case stress condition. In the figure below, it is shown how then such a reliability measurement is executed:

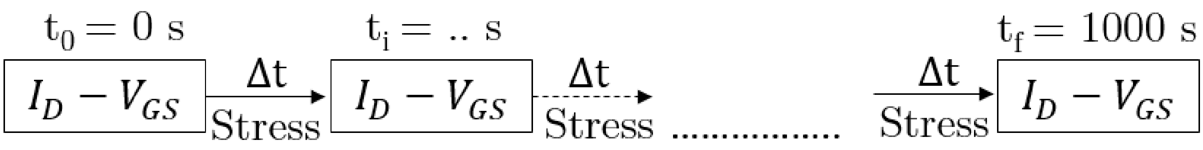

Figure 29: Schematic flowchart for reliability measurements.

First, an initial $I_D - V_{GS}$ measurement is performed, from which the $\mathrm{R_{ON}}$ and $V_{T,gm}$ are extracted. Subsequently, the worst-case stress condition is applied for a certain time Δt. After that, again the $\mathrm{R_{ON}}$ and $V_{T,gm}$ are extracted from a $I_D - V_{GS}$ measurement. This process is repeated for a desired number of cycles such that the degradation in $\mathrm{R_{ON}}$ and $V_{T,gm}$ can be mapped as a function of time. The total stress time used in this work is 1000 s.

4.2.3 Capacitance measurements

Determining the capacitances is done by measuring the AC impedance of the DUT. The working principle of an AC impedance meter (in this case the CMU), is shown in Figure 30. Out of the high current terminal (HCUR), an AC voltage is applied to the DUT. The current through the DUT is measured by the low current terminal (LCUR), and the voltage across the DUT is measured by the high and low potential terminals (HPOT and LPOT). The current and the voltage are measured in a phase-locked manner, such that the phase angle between them can be extracted. Then, by using the measured impedance amplitude ($|Z|$) and the phase angle (θ), the impedance Z can be determined. For the impedance of the DUT, there are two common models. The default model is a conductance (G) in parallel with a capacitor (Figure 30). Another option, is a resistance ($R = 1/G$) in series with a capacitor (Figure 30). Usually, the aim is to perform measurements where $\theta = \pm 90°$, such that the DUT can be considered as a capacitance only. However, when $\theta \neq \pm 90°$, it is important which model is used for interpreting the capacitance. When a parallel conductance becomes significant the default model should be used and when a series resistance becomes significant the other model should be used. In this work, the default model was used and the phase angle was observed to be near 90° at all times. [58]

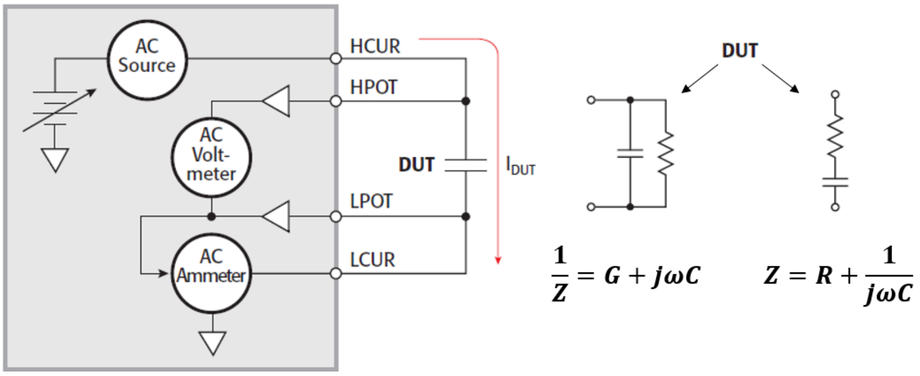

Figure 30: Left: Schematic representation of an AC impedance meter. Right: Modeling of the DUT, can be either a conductance in parallel with a capacitor (used in this work) or a resistance in series with a capacitor. [84]

The capacitance measurements that are done are: the total gate capacitance (C_{gg}), the gate-to-channel capacitance (C_{cg}), the drain-to-gate capacitance (C_{dg}) and the drain-to-source capacitance (C_{ds}). In the table below, it is shown for each measurement how the device's terminals are connected to the CMU (high=source, low=sensor). The frequency and the amplitude of the test signal are determined by trial-and-error to obtain a good signal-to-noise ratio. The specific settings will be indicated for each measurement. Moreover, before doing a measurement a calibration is performed to correct for attenuation and phase changes in the cables (open correction). Finally, it should be mentioned that for each measurement on a device also a so-called de-embedding structure is measured. These are structures on the wafer that do not contain the device but do have the bond pads and metal connections. By repeating the capacitance measurements on these, capacitances involved in the connection of the devices can be corrected for.

Table 3: Settings for the capacitance measurements. Gate voltage rating: 1.8V for GO1, 3.3V for GO2 and 5V for GO3 gate oxides. (S=Source, B=Body, D=Drain, G=Gate)

Measurement	Low	High	Ground	Voltage range
C_{gg}	G	S=B=D	-	±Gate voltage rating
C_{cg}	S=D	G	B	±Gate voltage rating
C_{gd}	G	D	S=B	0V - Voltage rating
C_{sd}	S=B	D	G	0V - Voltage rating

Chapter 5: Oxygen-inserted layer devices

Incorporating partial monolayers (PM's) of oxygen below the channel was proven to effectively enhance the channel mobility for standard CMOS devices (subsection 2.4.2) [2]–[7]. This is therefore potentially a promising way to reduce the $R_{ON}A$ for devices in NXP's technology platform without compromising the BV. To put this to the test, a feasibility study has been conducted in which this technique is applied to the full range of devices available in NXP's technology platform. Next to integrating such an oxygen-inserted layer (OIL), other process adjustments must be made to make the NXP devices compatible with this technique. The thermal budget must be reduced to prevent out diffusion of the oxygen (subsection 2.4.3) and the channel doping implants must be tuned to compensate for the decrease in V_T due to the super-steep-retrograde well (SSRW) doping profile (subsection 2.4.4). For this purpose, TCAD simulations are used to obtain a first order approximation for these process adjustments. However, due to the novel nature of this technique, the predictability of simulations is limited. Therefore, a split setup is used for the wafer lot. In this split setup, wafers are incorporated which have small changes in the processing with respect to the original flow (e.g. in the presence of an OIL, in thermal budgets and/or in doping doses) to get a clear overview of the influence of the adjustments. Once the wafers are processed, the factory performs a so-called PCM (Process Control Monitoring) evaluation. This PCM data already contains some electrical characteristics of a limited set of devices. This is therefore a good first indicator of the effect of the OIL. Any more detailed investigation on specific wafers is then performed at NXP.

In this chapter, a summary of the performed work concerning this approach to reduce the $R_{ON}A$-BV trade-off will be given without going into any detailed specifications and results for confidentiality reasons.

5.1 Summary

In order to obtain a first order approximation of the process adjustments that are needed to integrate the OIL correctly, simulations of the devices with and without the OIL are set up. The first aim of these simulations was to find new settings for the thermal budgets applied during processing for the devices with the OIL. The requirement for this was that the thermal budgets after the incorporation of the OIL should be reduced (subsection 2.4.3), while still maintaining enough overall thermal budget compared to the original process flow. The second aim of the simulations was to find new settings for the channel implant doping doses for the devices with the OIL. This is needed because the OIL effectively decreases the average doping below the gate interface, thereby lowering the V_T (Equation (2)). This should be corrected for since the devices are typically part of extensive electrical circuits where the V_T is to be fixed within in a certain range. Moreover, this also simplifies the comparison between the current characteristics of the original devices and devices with the OIL. So, based on the simulations, new (best-guess) settings of the process flow for the devices with the OIL are determined. Subsequently, a split could be set up for the wafer lot, in which the most significant split between the wafers is the presence of the OIL. Moreover, for both wafer types also small splits are incorporated in thermal budgets and channel implant doping doses because of the low predictability of the simulations.

On the wafers, three kind of device types were contained which are shown by simulation (n-type shown here) in the figure below. Note that here the OIL is not incorporated.

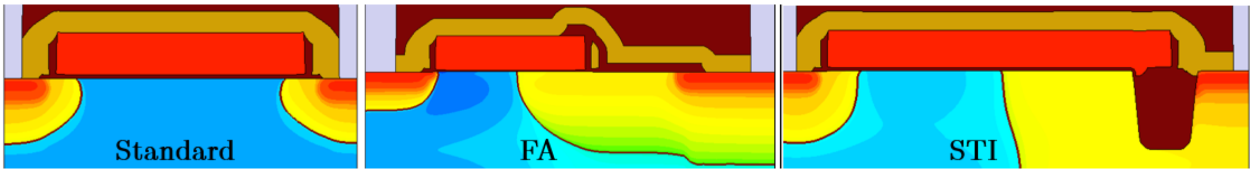

Figure 31: The three device types (n-type shown here) that are contained in the (oxygen-inserted layer) wafer lot. Here the color in the silicon indicates the doping concentration (red: n-type, blue: p-type). Left: Standard MOSFET (≤5V). Middle: Free-analog (FA) device (12V). Right: Shallow Trench Isolation (STI) device (16V).

In Figure 31 (left), a standard MOSFET is shown, which come in drain voltage ratings up to 5V. In Figure 31 (middle and right), two LDMOS device types are shown, the FA device (already shown in subsection 3.1.2) suited for 12V drain application and the STI device suited for 16V drain application. The STI device is a (semi-vertical) LDMOS design in which the gate is fully extended to the drain and a STI is situated below the gate at the drain side. This extended gate works as a FP (subsection 2.3.1), improving both the on-state and off-state characteristics. In addition, the STI makes part of the drift region vertical and optimizes the electric field distribution near the drain. The typical current paths for these devices in a RON measurement (V_{DS}=0.1V and full gate drive) are shown in the figure below.

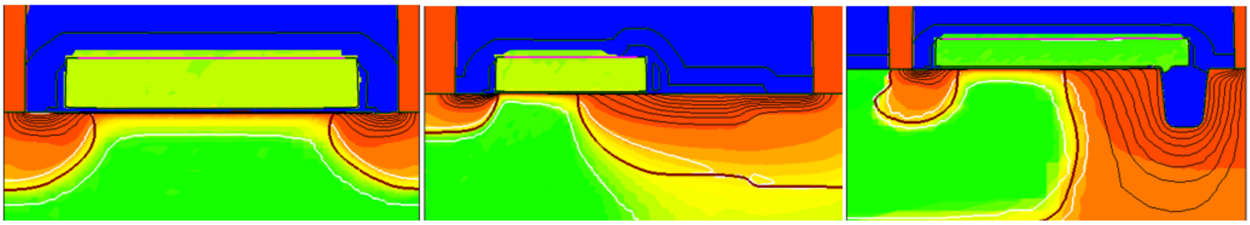

Figure 32: Absolute current density (towards red implies high current density whereas towards blue implies low current density) and current potential (black lines) for the standard MOSFET (left), FA device (middle) and STI device (right).

For the standard MOSFET (Figure 32 (left)), the current solely flows through the channel. Therefore, it is expected that integrating the OIL below the channel increases the linear drain current ($I_{D,lin}$) due to an increase in channel mobility [3]–[7]. In the drain-extended devices (Figure 32 (middle and right)), the current is solely at the Si-SiO$_2$ interface in the channel region, starts to diffuse into the bulk in the accumulation region and is mostly in the bulk for the drift region. Note that for the STI device, the drift region is along the STI edge. These current paths are the same as for the typical LDMOS device (Figure 20). For these type of devices, it was already discussed in subsection 2.4.6 that it is hard to form a hypothesis of the effect of the OIL. Experimental results are therefore needed to reveal the outcome.

After the wafer lot was produced and PCM data was made available, the improvement in $I_{D,lin}$ due to the OIL for the various devices was extracted and analyzed. Since this macroscopic parameter ($I_{D,lin}$) directly depends on the mobility, it is a good first indicator of the effect of the OIL. The device that showed the largest improvement is subsequently analyzed in more detail. First, the $I_{D,lin}$ improvement is evaluated for the various wafer split variations as well as for varying dimensions of the device. This yielded valuable information of the effect of the process adjustments and device dimensions on the performance of the OIL.

Then, to assess whether the improvement in $I_{D,lin}$ is indeed due to an improvement in the channel mobility, μ_{ch} as a function of E_{eff} (section 2.4) is extracted for several wafers with and without the OIL. This is done using the so-called CV-split method [85]. In this method, two measurements must be performed: $I_D - V_{GS}$ and $C_{cg} - V_{GS}$. The C_{cg} measurement can subsequently be used to obtain the inversion charge density ($|Q'_{inv}|$) as a function of V_{GS}, by integrating the $C_{cg} - V_{GS}$ curve. Then, using Equation (3) (where R_{ch}=V_{DS}/I_{DS}), μ_{ch} can be calculated for each V_{GS}. In addition, V_{GS} can be converted to E_{eff} using Equation (24). For this, the threshold voltages extracted from the $I_D - V_{GS}$ curves and the oxide thicknesses extracted from the $C_{cg} - V_{GS}$ curves are used. The resulting μ_{ch}-E_{eff} curves are then plotted together with the universal high field mobility [51] (for checking purposes) and compared to each other and to what is expected from literature [3]–[7].

In addition, in subsection 2.4.5 it was stated that due to the quasi-confinement effect not only the average transport mobility increases but also the gate leakage reduces by ~30-50% [5]. Therefore, a final asset that can be used to independently asses the quasi-confinement effect of the OIL, is the gate leakage. Gate leakage in a MOS structure is mainly due to quantum mechanical tunneling which decreases exponentially with increasing oxide thickness (t_{ox}) [38], [51]. Now, since t_{ox} typically slightly differs between wafers, the gate leakage of the devices with OIL cannot be compared one-to-one with the original devices. Luckily, there is always a small variation present in t_{ox} on the wafer (<1Å), such that the gate leakage dependency on t_{ox} can be extracted. This then requires two measurements at various locations on the wafer. First, the $C_{gg} - V_{GS}$ measurement, from which C_{ox} and t_{ox} can be extracted. Secondly, the gate current (I_G) at the maximum gate voltage and the source, drain and body at ground. These measurements are again done one several wafers with and without OIL. The resulting ln(I_G)- t_{ox} data is then linearly fitted for each wafer, and compared to each other and to what is expected from literature [5].

Lastly, to assess whether the OIL is integrated correctly and has the desired effect on the channel doping, TEM device cross-sections and SIMS profiles through the channel (of oxygen and boron) were requested for several wafers with and without OIL. Besides the visual inspection of the (influence of the) OIL, from these also valuable parameters could be extracted such as: the oxygen dose, channel doping concentration, cap layer thickness and OIL thickness. These parameters are subsequently compared to what was expected from simulations and based on that could be used to clarify some of the observations made in the electrical analysis. Moreover, additional supportive information was obtained by a comparison of the current experiment with those in literature [4]–[7], [70]. Nevertheless, the experiments conducted in this work concerning the OIL concept are not conclusive yet and follow-up work is therefore necessary.

Chapter 6:
Contact field plate devices

In 2017 it was shown by L. Wei et al. [8], that a contact module which is normally used for connecting the source, drain and body, can be used as a FP. This is simply done by placing the contact above the drift region, resulting in the so-called contact field plate (CFP). This was shown to mainly improve the reliability of the device by hot carrier injection (HCI) degradation, while no additional masks or process changes (i.e. costs) are needed [8]. Additionally, like all FP's, the CFP has also the potential to improve the $R_{ON}A$-BV trade-off (subsection 2.3.1).

One of the devices in NXP's technology platform, namely the Free-Analog (FA) device, is a typical LDMOS device that could benefit from this technique. In this work, it is explored by both experiment and simulation how the CFP affects its characteristics and how it can be used to optimize the $R_{ON}A$-BV trade-off. In order to do this, first TCAD simulations are set up for FA- and CFP devices. Some of the CFP-related critical dimensions for the simulations are obtained from TEM cross-sections. The simulations are subsequently used to support and substantiate electrical measurements. Electrical measurements are performed over a wide range of device variations in which important device parameters are varied, such that their dependencies on the $R_{ON}A$ and BV can be extracted. Moreover, these measurements can also be used to extract the independent resistance contributions and additionally serve as a benchmark for the proposed resistance separation method based on a single device. Next to the basic electrical measurements, also reliability and capacitance measurements are performed. These are important indicators for the lifetime and the switching losses, which may lay constraints on certain device parameters. Lastly, based on all the gathered information and dependencies, the CFP device will be optimized by simulation for the initial 12V application as well as for higher voltages and compared to the current industry standard.

In this chapter, first it will be discussed what the CFP devices look like, what are its most important parameters and what kind of variations are present in the wafer lot. Subsequently, the experimental results will be presented and where necessary supported by simulations. Lastly, it will be shown by simulation how the CFP device can be optimized.

6.1 Devices

The CFP devices are almost the same as the earlier mentioned FA devices (subsection 3.1.2 and section 5.1) with the 12V (on-state) drain rating and 5V gate rating (GO3 gate oxide), only now with an extra contact situated above the drift region (attached to the gate) acting as a FP. In the figure below, the result of a process simulation of such a CFP device (with default dimensions) is shown. In this figure, also the most important dimensions and doping implants are indicated. It should be mentioned that the p$^+$ body contact is not simulated here to ease the simulations. Instead, the body contact is defined at the edge of the deep pwell implant which is meant for low ohmic body connection (see also Figure 22p).

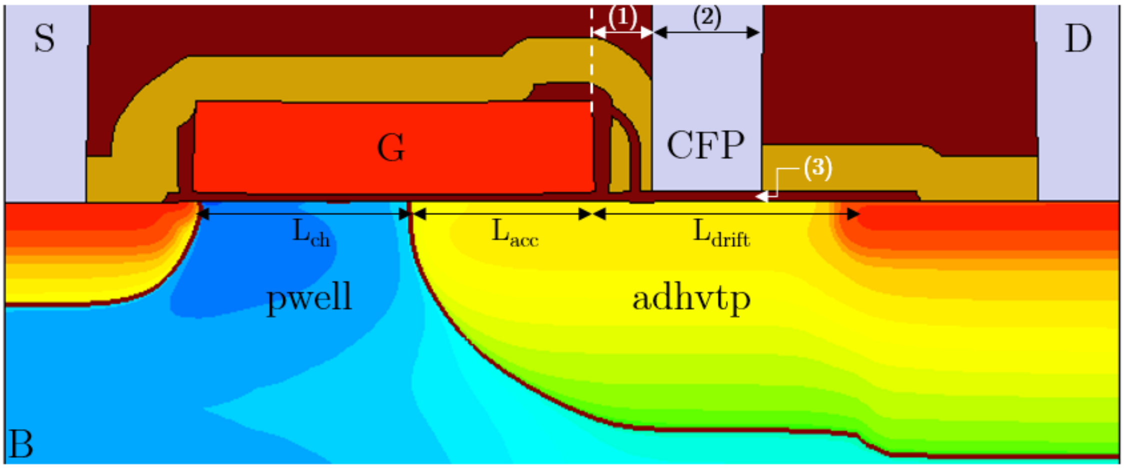

Figure 33: Simulated n-type CFP device with GO3 gate oxide. Here the color in the silicon indicates the doping concentration (red: n-type, blue: p-type). The most important doping implants and dimensions are indicated. ((1) = poly-CFP spacing, (2) = CFP length and (3) = $t_{ox,FP}$). The pitch of the default device is 2.40 µm.

The doping implants in respectively the channel and drain-extension are the so-called pwell- and adhvtp-implants. The L_{ch}, L_{acc} and L_{drift} dimensions have the same meaning as for the standard LDMOS device (Figure 5 (left)). For the CFP devices, there are also three other important dimensions denoted by the numbers 1-3. These are respectively the spacing from the polysilicon gate to the CFP (1), the length of the CFP (2) and the thickness of the oxide below the CFP ($t_{ox,FP}$) (3). In order to have an estimate of these dimensions, as well as to check whether the CFP is integrated nicely, TEM cross-sections are obtained from the default FA- and CFP device. In total, there are cross-sections made of two two-fold structures (one at the center and one at the edge of the wafer), such that cross-sections of four devices are available. In the figure below, one cross-section for each device is shown. At first sight, it seems that the simulations (Figure 31 (left) and Figure 33) closely resemble the actual devices and that the CFP is integrated nicely.

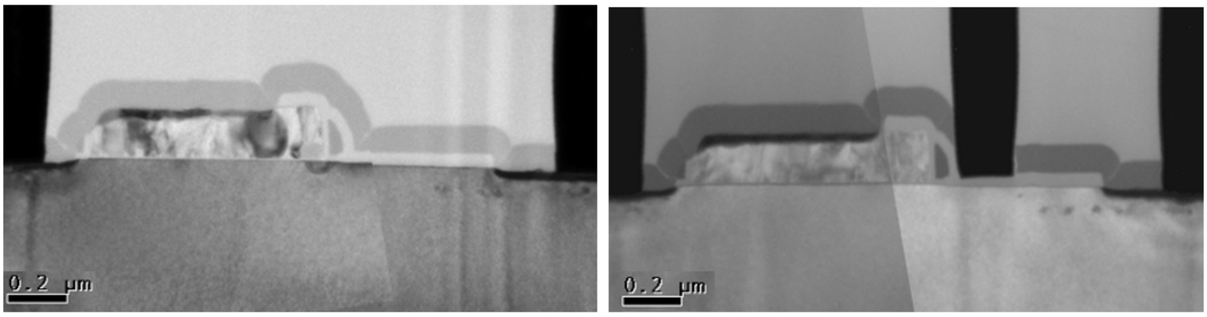

Figure 34: TEM cross-sections of the default FA device (left) and the default CFP device (right). Both cross-sections look very similar to the simulations (Figure 31 (middle) and Figure 33) and the CFP seems to be integrated nicely.

From TEM cross-sections with higher magnification, the three critical dimensions are extracted and compared to those extracted from an initial process simulation (see Table 4). Relatively large differences are present, indicating the need for calibration of the process simulation. For this, the TEM image is used that showed the worst-case dimensions (for off-state breakdown). This is because in practice the location along the width of the device where the dimensions are least optimal, will always initiate the breakdown of the whole device.

Table 4: Important CFP-related dimensions extracted from four TEM cross-sections (both the average and worst-case for off-state breakdown) and from simulations before (initial) and after calibration with respect to the worst-case dimensions. All dimensions are in nm.

Dimension	**Simulation (before calibration)**	**TEM (average)**	**TEM (worst-case)**	**Simulation (after calibration)**
Poly-CFP	134	104	107	107
CFP length	143	194	200	200
$t_{ox,FP}$	40	23	21	21

Just as in the OIL wafer lot, the CFP wafer lot has a split in the wafers and in the devices on the wafers (for both the FA- and CFP devices). Between the wafers, the thickness of the SiProt layer (subsection 3.1.2) is varied to obtain various $t_{ox,FP}$ thicknesses relevant for the CFP devices. The variations in the SiProt layer thickness are 55, 68 (default) and 85 nm, which based on the (worst-case) TEM cross-section of the default CFP device result in $t_{ox,FP}$ thicknesses that are approximately 47 nm thinner. Between the devices on the wafer, the relevant lateral dimensions (L_{ch}, L_{acc} and L_{drift}) are varied. In the table below, an overview of the most important variations on this wafer lot are shown.

Table 5: Overview of the most important dimension variations (in µm) in the wafer lot for the FA- and CFP devices. Also, the default settings are indicated. The pitch of the default FA- and CFP device is 2.40 µm.

	L_{ch} (default = 0.35)	L_{acc} (default = 0.30)	L_{drift} (default = 0.40)	$t_{ox,FP}$ (default = 0.021)
FA devices	0.20 - 0.50	0.10 - 0.40	0.32 - 0.50	-
CFP devices	0.35	0.10 - 0.40	0.32 - 0.50	0.008 - 0.380

Now that it is shown what the CFP devices look like and what the important parameters and variations are, the results obtained from both experiments and simulations will be discussed.

6.2 Results

In this section, first it will be shown what is the effect of just adding the CFP to the default FA device on the on- and off-state characteristics by both measurements and simulations. Thereafter, R_{ON}A-BV dependencies on important dimensions are shown and supported with simulations. These dependencies are subsequently used to extract the independent resistance contributions and in turn used as a benchmark for a proposed resistance separation method based on a single device. Lastly, the influence of the CFP on the lifetime and power losses will be assessed based on reliability and capacitance measurements.

6.2.1 Basic electrical measurements

In Figure 35 (left), the $I_D - V_{GS}$ curves for the default FA- and CFP device are shown (obtained from measurements and simulations):

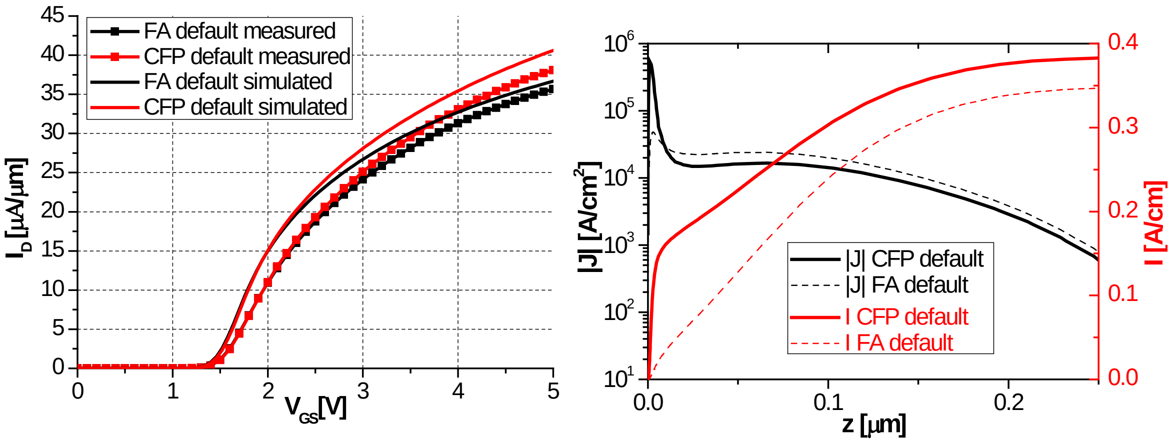

Figure 35: Left: Measured and simulated $I_D - V_{GS}$ curves for the default FA- and CFP device. The linear drain current is increased for the CFP device (from both simulation and measurement) due to the additional accumulation. Right: Simulated current density (|J|) and integrated current (I) as a function of the distance to the Si-SiO₂ interface below the location of the CFP for the default FA- and CFP device at V_{GS}=5V and V_{DS}=0.1V. The additional accumulation at the interface results in an increase of the current

From Figure 35 (left), it can first of all be seen that the V_T is matched well between the simulations and measurements (both 1.45V). The linear drain current (at V_{GS}=5V) is slightly overestimated by the simulation (~3% for FA device and ~7% for CFP device). This is likely due to metal and contact resistances that are not incorporated in the simulation. In addition, the reason for the slightly higher overestimation for the CFP device is probably the result of process variation of the CFP-related dimensions over the wafer (the measured device is not on the same location on the wafer as the device that is used for calibration). Nevertheless, the improvement in $I_{D,lin}$ due to the accumulation below the CFP can be observed from the measurements (~7%) as well as from the simulations (~11%). In Figure 35 (right), this additional accumulation can clearly be seen from the simulated current density (|J|) and integrated current (I) as a function of the distance to the Si-SiO₂ interface below the location of the CFP. These curves are extracted from the simulations shown in Figure 69 (Appendix C).

In Figure 36 (left), the off-state breakdown curves for the default FA- and CFP device are shown (obtained from measurements and simulations):

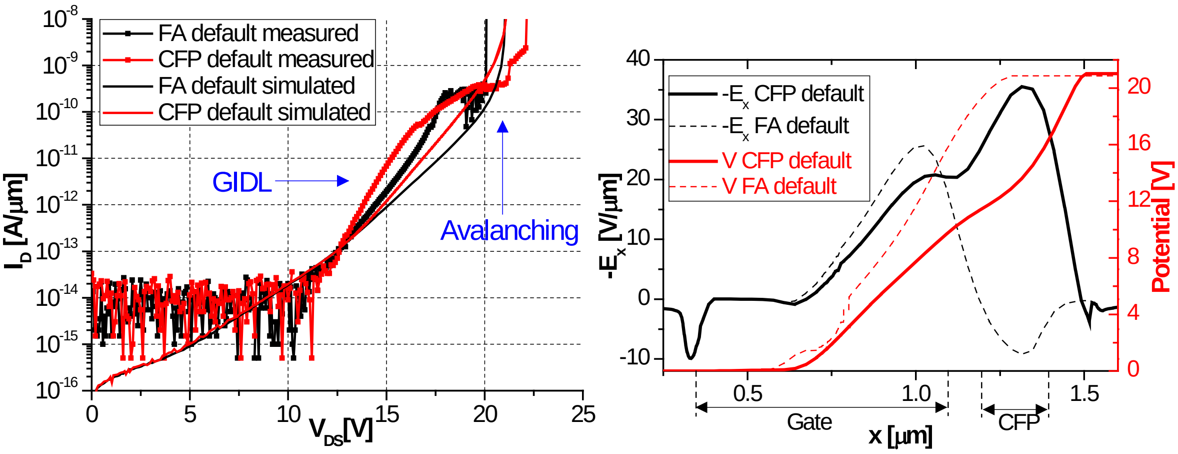

Figure 36: Left: Measured and simulated off-state breakdown curves for the default FA- and CFP device. From the simulations there is basically no improvement in BV. From the measurements, a ~2V improvement can be observed, however may be a result of walkout [86]. *Right: Simulated lateral electric field and potential through the default FA- and CFP device (off-state breakdown simulation at I_D=1x10⁻⁸ A/µm). Adding the CFP extends the depletion region but also introduces a second more dominant field peak, resulting in no improvement in BV.*

From Figure 36 (left), it first of all stands out that the simulations and measurements differ in the types of curves. In the simulations, the slope gradually increases as V_{DS} increases due to GIDL (subsection 2.1.3) before the avalanche breakdown sets in. In the measurements, also first some soft (gradual) breakdown is observed due to GIDL, however then converges to a plateau before it breaks down due to avalanching. This deviation from the simulation is suggested to be contributed to so-called walkout, which is a term that is used to describe the upward shift of the breakdown voltage with increasing current. This is a result of charge injection into the oxide which optimizes the local electric field distribution and thereby increases the BV [86]. From the simulated curves, it is evident that there is basically no improvement in the BV between the FA- and CFP device. The reason for this can be seen in Figure 36 (right), where the simulated lateral electric field and potential through the lateral direction (x) of the devices are shown (at breakdown). From this figure, it is evident that the CFP extends the depletion region as was expected (Figure 9). However, instead of lowering the lateral electric field, a second (more dominant) field peak below the CFP is introduced. In the figure below, it can be seen from the cross-sections belonging to these breakdown simulations that this effectively concentrates the center of impact ionization (II) more towards the drain, which results in no improvement in the BV.

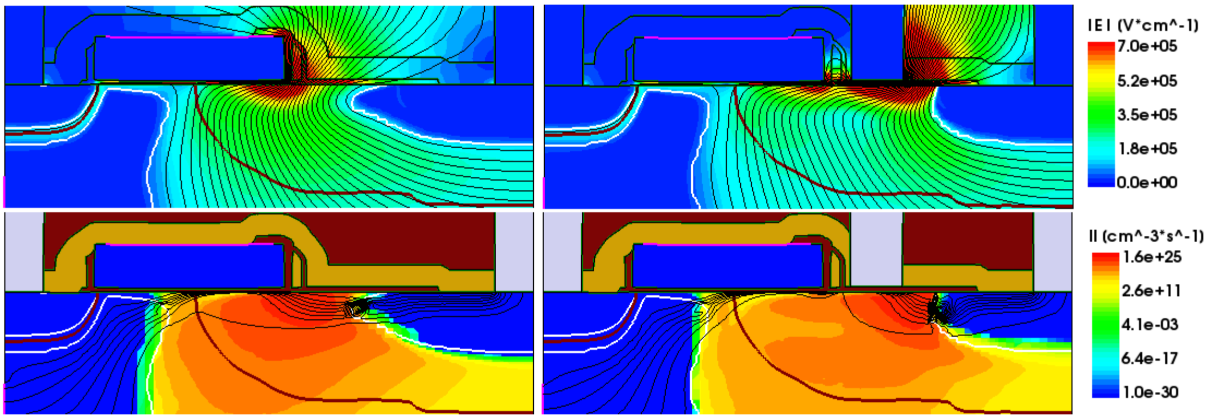

Figure 37: Off-state breakdown simulations of the default FA- (left) and CFP device (right) at I_D=1x10^{-8} A/µm. Top: Magnitude of electric field (|E|) and equipotential lines (black lines). Bottom: Impact ionization (II) and current potential lines (black lines). The CFP helps to deplete the drift region, but also induces a large electric field peak at its edge. The center of impact ionization is therefore concentrated more towards the drain. As a result, there is basically no effect on the BV.

To compare the simulated BV's with the measured BV's, it is not reasonable to do this at the current where the device actually breaks down (~10 nA/µm) because charge injection is not simulated. Instead, for this purpose, the breakdown for these devices is defined as the onset of leakage at 1 pA/ µm in the rest of this section. At this current, it can be seen from Figure 36 (left) that the BV is slightly overestimated by the simulations (~5% for FA device and ~7% for CFP device). Moreover, adding the CFP actually reduces the BV by ~2% from the simulations and by ~4% from the measurements.

Clearly, the default CFP device is not optimized. An insightful way to show is, is shown in Figure 38 (left). Here, the BV is plotted as a function of the FP bias obtained from simulations and from measurements on a device with a separate bond pad for the CFP.

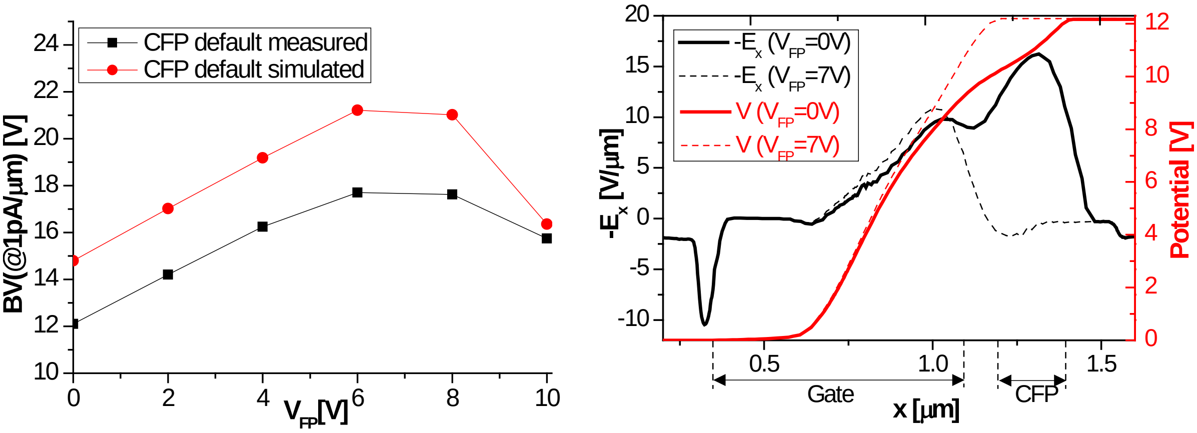

Figure 38: Left: BV as function of the FP bias from simulations and from measurements on a device with a separate CFP bond pad. The maximum BV is obtained for a FP bias of ~7V. Right: Simulated lateral electric field and potential below the gate and CFP for FP biases of respectively 0V and 7V for V_{DS} ≈12V. Applying 7V to the FP lowers the field peak at the CFP edge at the cost of the field peak at the gate edge and reduces the depletion width. This increases the BV by reducing the band-to-band generation as well as the impact ionization.

Although the simulations overestimate the BV, the same trend is observed as in the measurements where the BV peaks at approximately V_{FP} ≈7V (Figure 38 (left)). The optimum for a positive FP bias basically means that the CFP depletes the underlying silicon excessively for the default dimensions. The effect of this is shown in Figure 38 (right), where the lateral field and potential below the gate and CFP are shown from simulations at V_{FP}=0V and V_{FP}=7V for V_{DS} ≈12V. Here it can be seen that applying the positive V_{FP} lowers the field peak at the CFP edge at the cost of the field peak at the gate edge and reduces the depletion width. This decreases the leakage current by both reducing the band-to-band generation as well as the impact ionization (Figure 70 in Appendix C)

Next to the off-state breakdown, it is also important to have a look at the on-state breakdown. In Figure 39, the $I_D - V_{DS}$ curves for three different gate voltages are shown obtained from both measurements and simulations. Note that the measurements are only performed until 12V (voltage rating), this is because thermal heating (i.e. melting) destroys the devices for higher drain biases due to the high current density (as already can be seen for the CFP device at full gate drive). Also, because self-heating is not simulated, the saturation currents observed in the measurements are significantly lower than in the simulations. It should be mentioned that for devices intended for switching application, typically the AC on-state breakdown curves are of interest (instead of DC) which can be obtained by transmission-line pulse measurements [58]. In these, the thermal heating has no/less effect due to the short nature of the pulses, thereby allowing to trace the full on-state breakdown curves as simulated.

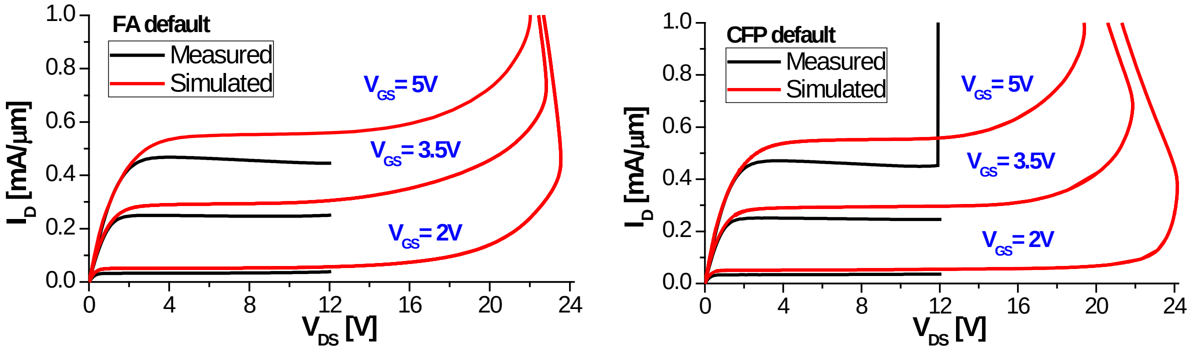

Figure 39: Measured and simulated $I_D - V_{DS}$ curves for three different V_{GS} for the default FA device (left) and default CFP device (right). Note that the measurements are only performed until 12V, this is because self-heating destroys the devices for higher drain biases (as can be seen for the CFP device at full gate drive). Self-heating is not simulated resulting in higher saturation currents than in the measurements. From the simulations, adding the CFP increases BV_{ON} for the lower gate voltages whereas it decreases the BV_{ON} for full gate drive.

Nevertheless, several observations can be made from Figure 39. From both the measurements and simulations, it can be seen that the saturation currents are approximately equal for the default FA- and CFP device. This is because for increasing drain bias, the benefit of accumulation below the CFP vanishes since the potential below the CFP becomes higher than the FP bias. In fact, the CFP then depletes the underlying silicon thereby bending the current route further into the bulk. As the doping concentration in the drain-extension is reasonably uniform, this has no significant effect on the current. This effect can clearly be seen in the top figures of Figure 40 and Figure 41, where the simulated current density and current potential lines are shown through the devices at V_{DS} ≈12V for V_{GS}=2V and V_{GS}=5V. Moreover, from the simulated curves, it can be observed that the CFP increases BV_{ON} for the lower gate voltages whereas it decreases the BV_{ON} for full gate drive. The reason for this is evident from the electric field and impact ionization cross-sections in Figure 40 and Figure 41. For low gate voltage (Figure 40), adding the CFP results in two more moderate field peaks instead of the strong field peak at the gate edge, thereby reducing the impact ionization. For large gate voltage on the other hand (Figure 41), adding the CFP displaces the field peak at the gate edge to the CFP edge where it becomes stronger, resulting in more impact ionization. Although BV_{ON} is not the main parameter of interest to optimize in this work, it is still important to keep in mind. In the next subsection, the focus will again be on the $R_{ON}A$ and BV and in particular how they depend on the various dimension variations.

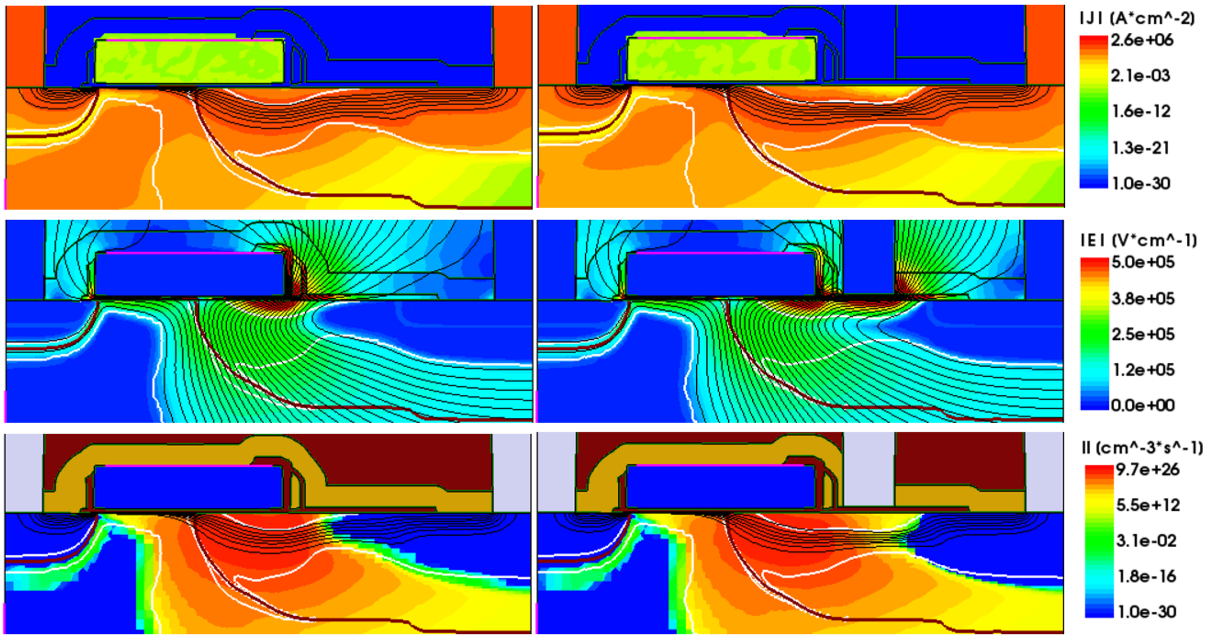

Figure 40: On-state simulations at V_{DS} ≈12V and V_{GS} =2V for the default FA device (left) and default CFP device (right). Top: Magnitude of current density (|J|) and current potential lines (black lines). Middle: Magnitude of electric field (|E|) and potential lines (black lines). Bottom: Impact ionization (II) and current potential lines (black lines). The CFP depletes the underlying silicon which bends the current route more into the bulk. Moreover, this results in two more moderate field peaks instead of the strong field peak at the gate edge, thereby reducing the impact ionization and hence increasing BV_{ON}.

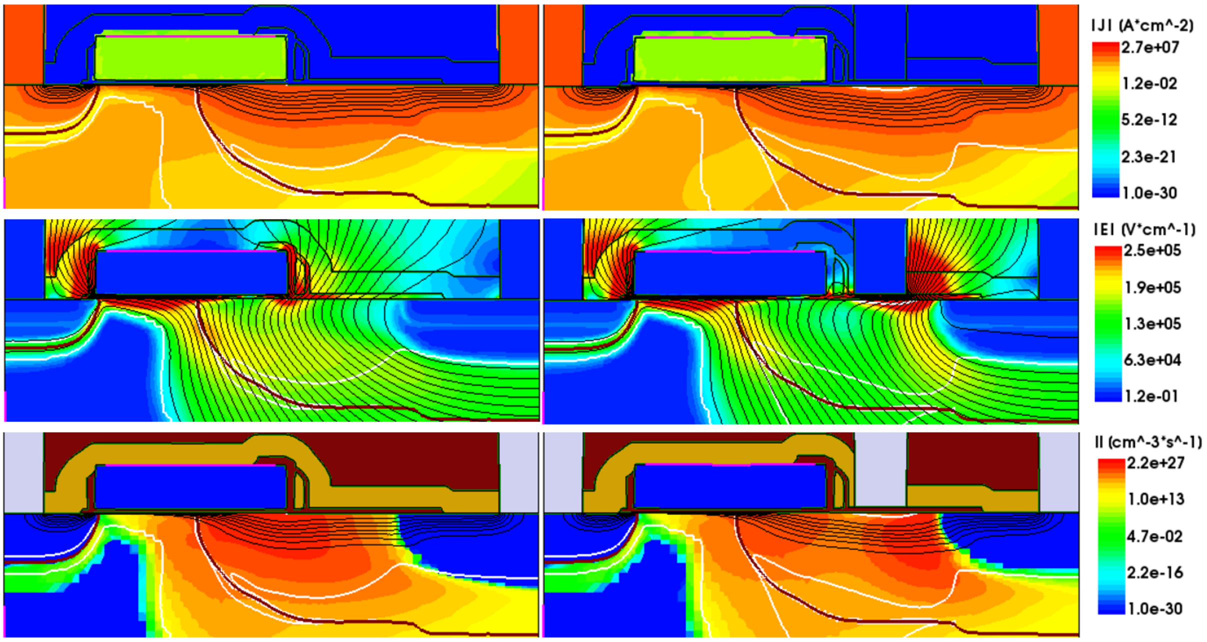

Figure 41: On-state simulations at $V_{DS} \approx 12V$ and $V_{GS} = 5V$ for the default FA device (left) and default CFP device (right). Top: Magnitude of current density (|J|) and current potential lines (black lines). Middle: Magnitude of electric field (|E|) and potential lines (black lines). Bottom: Impact ionization (II) and current potential lines (black lines). The CFP depletes the underlying silicon which bends the current route more into the bulk (although less than for $V_{GS} = 2V$). Moreover, this displaces the field peak at the gate edge to the CFP edge where it becomes stronger, resulting in more impact ionization and hence reducing BV_{ON}.

6.2.2 Electrical dependencies on dimensions

In the figure below, the dependencies of R_{ON}A and BV on L_{drift} obtained from measurements and simulations are shown for both the FA- and CFP devices.

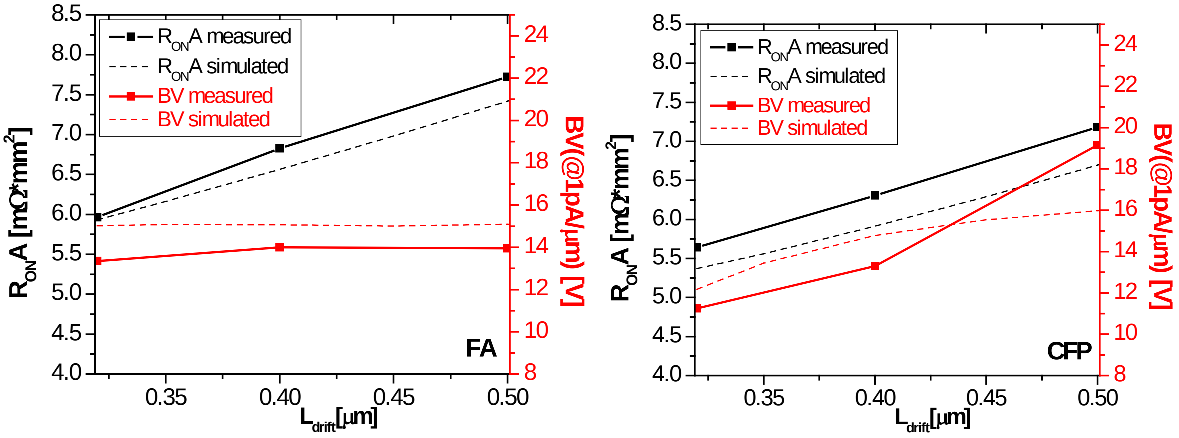

Figure 42: Dependencies of $R_{ON}A$ and BV on L_{drift} obtained from both measurements and simulations for the FA- (left) and CFP devices (right). Adding the CFP provides an effective way to scale to higher BV's.

From Figure 42, it can be seen that for both device types $R_{ON}A$ increases with L_{drift}, as could be expected from Equation (10). The BV however, stays approximately constant for the FA devices, whereas it increases with increasing L_{drift} for the CFP devices. This is because as L_{drift} is increased, the depletion width (W_D) is basically unaffected for the FA device since the depletion only occurs from the body. While for the CFP device, the additional depletion from the CFP results in a larger W_D for increasing L_{drift}

(up to a certain L_{drift}). This effect is illustrated in Figure 71 (Appendix C). Clearly, the CFP provides an effective way to scale to higher BV's.

In the figure below, the dependencies are shown as a function of L_{acc}.

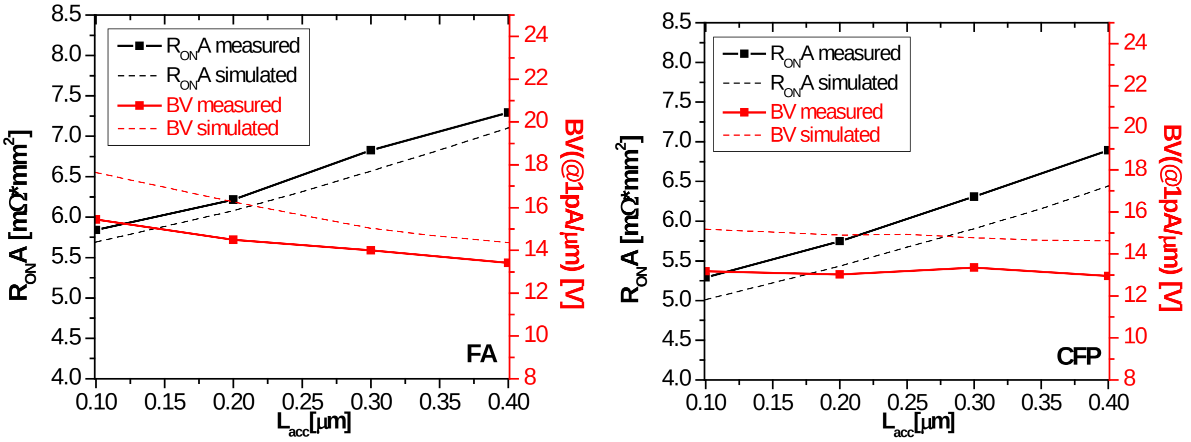

Figure 43: Dependencies of $\mathrm{R_{ON}A}$ and BV on L_{acc} obtained from both measurements and simulations for the FA- (left) and CFP devices (right). Adding the CFP lowers the BV for small L_{acc}.

From Figure 43, it can be seen that for both device types $\mathrm{R_{ON}A}$ increases with L_{acc}, as could be expected from Equation (9). The BV slightly decreases with increasing L_{acc} for the FA devices, whereas it remains almost constant for the CFP devices. This is because for increasing L_{acc} the field peak at the gate edge is increased/broadened, leading to more band-to-band tunneling (GIDL). For the CFP devices, this effect is less prominent since the field peak below the CFP is dominant. In Figure 72 (Appendix C), this is further illustrated. So, purely based on the $\mathrm{R_{ON}A}$ and BV, L_{acc} should be as short as possible. However, as will be shown in subsection 6.2.4, sufficient L_{acc} is needed to provide a reliable device.

In the figure below, the dependencies are shown as a function of L_{ch}. Note that for the CFP devices no variations in L_{ch} are available on the wafers, such that only simulated data is shown.

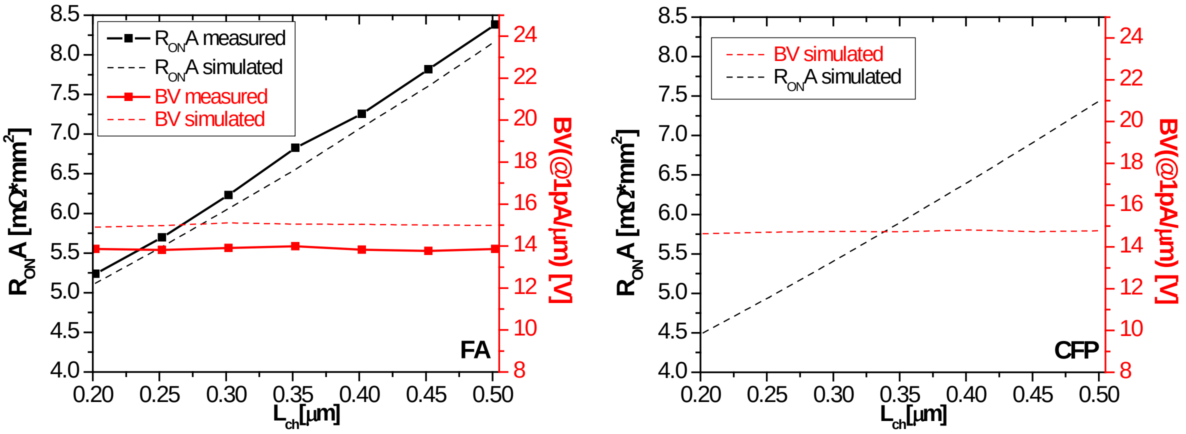

Figure 44: Dependencies of $\mathrm{R_{ON}A}$ and BV on L_{ch} obtained from both measurements and simulations for the FA devices (left) and only from simulations for the CFP devices (right). Adding the CFP has no result on the $\mathrm{R_{ON}A}$ and BV trends.

From Figure 44, it can be seen that for both device types that $\mathrm{R_{ON}A}$ increases (Equation (7)) and that BV remains constant with increasing L_{ch}, as expected. Therefore, decreasing L_{ch} is an effective way to improve the $\mathrm{R_{ON}A}$-BV trade-off. One drawback of this however, is the increase in V_T variability because of so-called V_T roll-off, which is a term to describe the decrease in V_T for decreasing L_{ch}. The reason for this is made evident in Figure 45 (left), where the simulated conduction band energy through the CFP devices (at $\mathrm{V_{DS}} = 0.1\mathrm{V}$ and $\mathrm{V_{GS}} = 0\mathrm{V}$) with L_{ch}=0.35 µm and 0.20 µm is shown. Here it can be seen that the energy barrier in the channel for the shorter L_{ch} is reduced, which is due to the increased influence of the source and drain(-extension) on the energy band diagram in the channel [51]. As a consequence, a smaller V_{GS} is needed to pull the barrier down to reach the inversion condition [51]. In Figure 45 (right), the V_T corresponding to the simulated data in Figure 44 is shown. Here it can be seen that for $L_{ch} < 0.35$ µm the V_T starts to decrease such that it becomes more sensitive to process variation in L_{ch}. This proces variation can be as large as 0.1 µm, such that this roll-off could be an issue for smaller channel lengths.

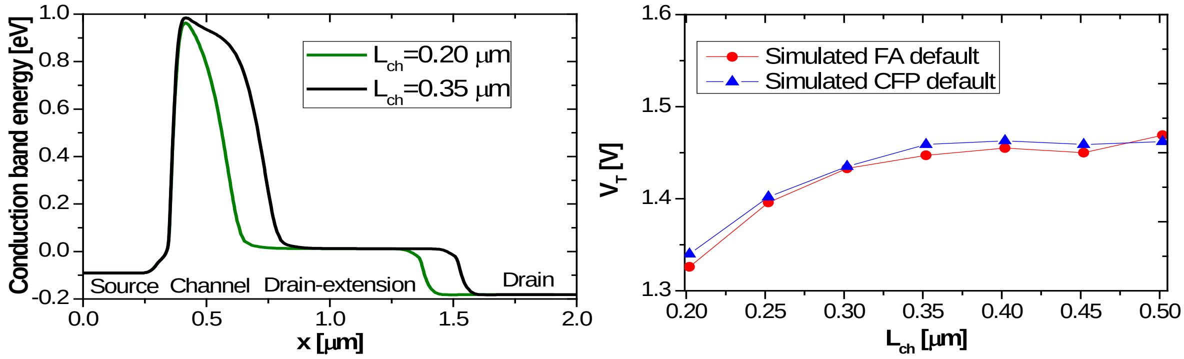

Figure 45: Left: Conduction band energy through the CFP devices with L_{ch}=0.35 and 0.20 µm (at $V_{DS} = 0.1V$ and $V_{GS} = 0V$). The energy barrier in the channel is reduced for L_{ch}=0.20 µm, resulting in lower V_T [51]. Right: V_T roll-off from simulation for the default FA- and CFP device. The decrease in V_T for L_{ch}<0.35 µm is evident.

The last dependencies that will be shown are of $t_{ox,FP}$, which are only relevant for the CFP devices. These dependencies are shown in Figure 46 (left). Note that the measurements are obtained from three different wafers.

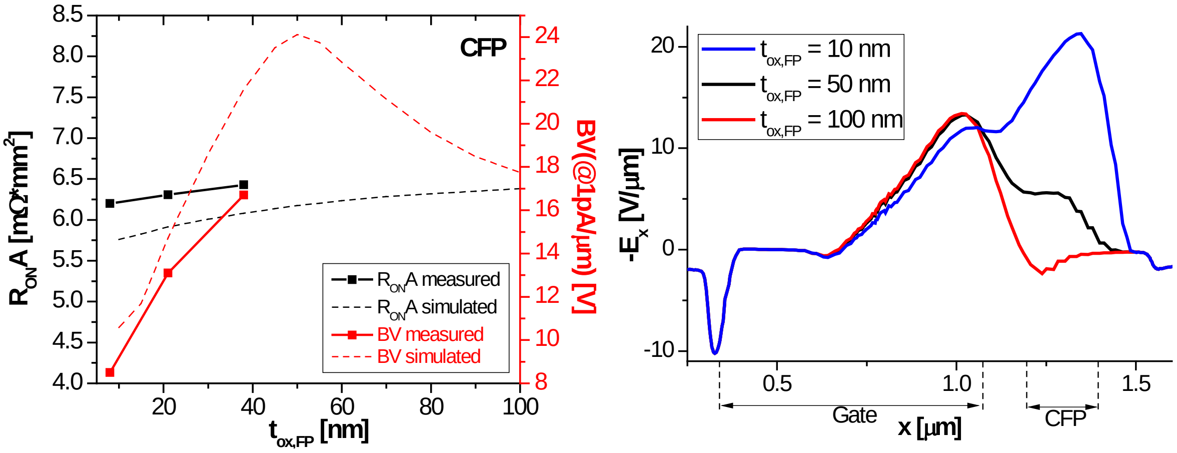

Figure 46: Left: Dependencies of $R_{ON}A$ and BV on $t_{ox,FP}$ obtained from both measurements and simulations of the CFP devices. A clear optimum in the BV can be observed for 50 nm, which is above the $t_{ox,FP}$ available on the wafers. Right: Simulated lateral electric field (at $V_{DS} \approx 12V$) below the gate and CFP for a $t_{ox,FP}$ of respectively 10, 50 and 100 nm. Below the optimum $t_{ox,FP}$ (50 nm) the field peak below the CFP becomes dominant, whereas above the optimum the field peak at the gate edge becomes dominant.

From Figure 46 (left), it can be seen that the RONA slightly decreases with decreasing $t_{ox,FP}$ due to the increasing accumulation below the CFP. The BV seems to have a clear optimum from the simulation around 50 nm. This is basically equivalent with the optimum in BV for positive V_{FP} (Figure 38 (left)), since applying a positive V_{FP} lowers the vertical field just as increasing $t_{ox,FP}$. Below this optimum the field peak below the CFP becomes dominant, whereas above this optimum the field peak at the gate edge becomes dominant (Figure 46 (right)). The maximal $t_{ox,FP}$ available (~ 38 nm) on the wafers still seems to be below this optimum. Therefore, the BV can significantly be increased by increasing $t_{ox,FP}$ while the RONA only slightly increases.

Besides examining how device parameters affect the RONA, it is also useful to determine the relative contributions of the series resistances to the total RON. This can for example be used to identify which contributions should be optimized or how adding the CFP affects the independent resistance contributions.

6.2.3 Resistance contributions

In this subsection, two methods are used to determine the different contributions to the RON of the default FA- and CFP device. The less accurate method utilizes only one device architecture and is able to differentiate between $R_{sd} + R_{drift}$, R_{ch} and R_{acc}. Whereas the more accurate method can differentiate between all resistance components, however needs device architectures with varying L_{drift}, L_{acc} and L_{ch}.

The first method makes use of $I_D - V_{GS}$ curves of the same device with different threshold voltages. These are obtained by applying varying source-to-body biases (V_{SB}), in this case -0.5, 0 and 0.5V. The curves are subsequently converted to RON$-V_{GS}$ curves by dividing the applied V_{DS} voltage (0.1V) through the drain current I_D. Then, the resulting curves are fitted to the LDMOS resistance model (subsection 2.2.1) for $V_{GS} > V_T + 0.5$. The parameters resulting from this fit are used to determine $R_{sd} + R_{drift}$, R_{ch} and R_{acc}. In Figure 47 (left), this approach is shown for the default FA device. The V_{FB} = -0.15V corresponds to a donor concentration of 10^{17} cm^{-3} in the accumulation region (Appendix A.1). This approach, to differentiate the resistance contributions, is clarified in more detail in Appendix A.2.

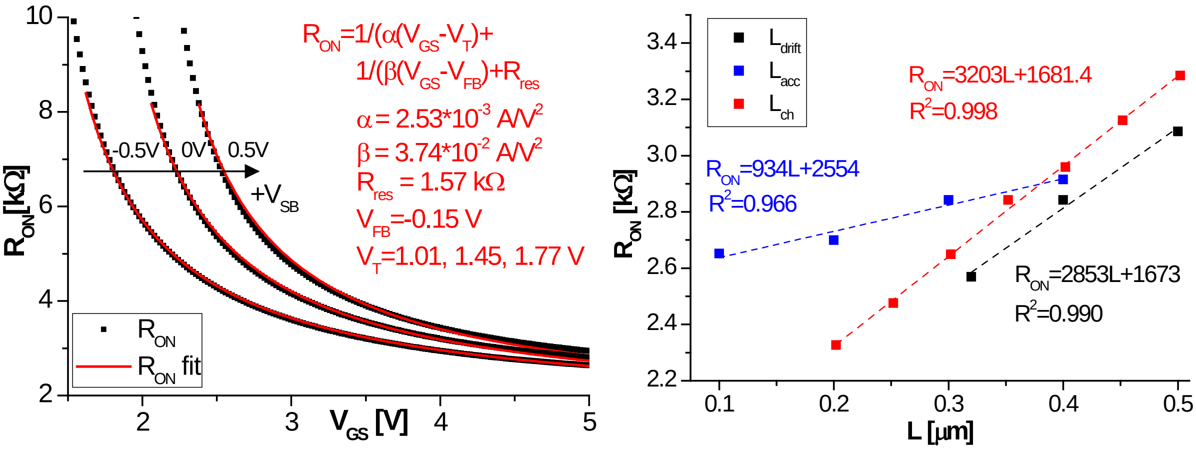

Figure 47: Extracting the series resistance contributions to the RON by two different methods for the default FA device. Left: The RON $-V_{GS}$ curves of a single device obtained at V_{SB}=-0.5, 0, 0.5V are fitted to the LDMOS resistance model. A more detailed description of this approach can be found in Appendix A.2.
Right: The RON dependencies on the L_{drift}, L_{acc} and L_{ch} dimensions obtained from multiple device architectures are fitted to linear equations and used to solve for each resistance contribution.

The second method makes use of the RON dependencies on the L_{drift}, L_{acc} and L_{ch} dimensions and employs the fact that the R_{drift}, R_{ch} and R_{acc} contributions scale linearly with these dimensions (subsection 2.2.1). By fitting the dependencies to linear equations, three equations can be obtained for the y-axis intercepts. This is shown for the default FA device in Figure 47 (right). For example, the y-axis intercept from the linear fit through the RON - L_{ch} data represents $R_{drift} + R_{acc} + R_{sd}$. In a similar way, equations can be obtained for $R_{ch} + R_{acc} + R_{sd}$ and $R_{drift} + R_{ch} + R_{sd}$ from the L_{drift} and L_{acc} dependencies respectively. The fourth and last equation needed for solving the system of equations, can simply be obtained from the total RON yielding an expression for $R_{ch} + R_{acc} + R_{drift} + R_{sd}$. It must be mentioned that since for the CFP devices no L_{ch} dependency is available, the R_{sd} obtained from the FA device is used to eliminate one equation.

In Table 6, the relative resistance contributions to the R_{ON} are shown for the default FA- and CFP device obtained from both methods.

Table 6: Relative resistance contributions to the R_{ON} for the default FA- and CFP device, obtained from two methods. From both methods, it is evident that adding the CFP results in a relative decrease of R_{drift} mainly at the expense of R_{acc}. Also, it can be seen that R_{ch} has a relatively large contribution, indicating room for improvement there.

	FA device		CFP device	
	Method 1	Method 2	Method 1	Method 2
R_{ch}	40.7%	40.9%	42.7%	41.2%
R_{acc}	1.9%	10.5%	4.3%	13.2%
R_{drift}	57.4%	41.2%	53.0%	37.3%
R_{sd}		7.5%		8.1%

From Table 6, several things can be observed about the differences between the methods, the influence of the CFP and the resistance contributions in general. First of all, it can be seen that the largest variation between the methods occurs for R_{acc}. This is probably because the effective length of the accumulation layer, in which carriers solely flow at the Si-SiO_2 interface, is less than the defined L_{acc} (Figure 5 (left)) [52]. This is a result of that the current starts to diffuse into the bulk in the accumulation region, which for example can be seen in Figure 69 (Appendix C). Therefore, the R_{acc} determined by method 2 represents the resistance of the whole accumulation region as defined. Whereas for method 1, R_{acc} only represent the resistance in the accumulation layer and the resistance part of the bulk current in the accumulation region adds to R_{drift}. This results in that R_{acc} determined by method 1 is smaller than for method 2 at the increase of R_{drift}. Secondly, from both methods it is evident that adding the CFP results in a reduction of the relative contribution of R_{drift}, mainly at the expense of R_{acc}. This was also expected due to the additional accumulation below the CFP. Lastly, mostly from method 2, it can be seen that R_{ch} and R_{drift} both approximately make up 40% and the R_{acc} and R_{sd} both approximately make up 10%. The relatively high contribution from R_{ch} (without contributing to the BV), indicates that there is much room for improvement regarding the channel. The most effective/straightforward way this can be done is by reducing L_{ch}, which already became evident from Figure 44.

Next to that reducing L_{ch} is an effective way to reduce the $R_{ON}A$ without comprising the BV, this also applies for reducing L_{acc} (Figure 43). However, a constraint on this dimension exists in terms of reliability. How this is affected by the CFP will be discussed next.

6.2.4 Reliability

For the FA device, the critical parameter for hot carrier injection (HCI) is believed to be L_{acc}. For this device in the on-state, the electrical field is typically the highest at the Si-SiO_2 interface at the edge of the gate at the drain site. The high impact ionization in this region results in much scattering and thereby in HCI into the oxide. Therefore, the HCI increases with the number of charge carriers crossing the vicinity of this critical point. When L_{acc} is small, the current diffusion into the bulk is low, such that a large fraction of the carriers crosses the vicinity of this critical spot. When L_{acc} becomes larger however, a larger part of the current diffuses into the bulk, thereby decreasing the HCI. In Figure 48, the enhanced diffusion into bulk for increasing L_{acc} is illustrated by on-state simulations (at worst-case stress conditions in terms of reliability) of the FA- and CFP device with a L_{acc} of 0.2 µm and 0.3 µm. In addition, it should be mentioned that for a given amount of HCI, the injected charge in the oxide also has less influence on the current for increasing L_{acc}. On one hand, this because the injected charge is distributed over a larger

L_{acc}, such that the oxide charge density becomes lower. On the other hand, as the current flows deeper, the interaction between the fixed oxide charge and the charge carriers is reduced.

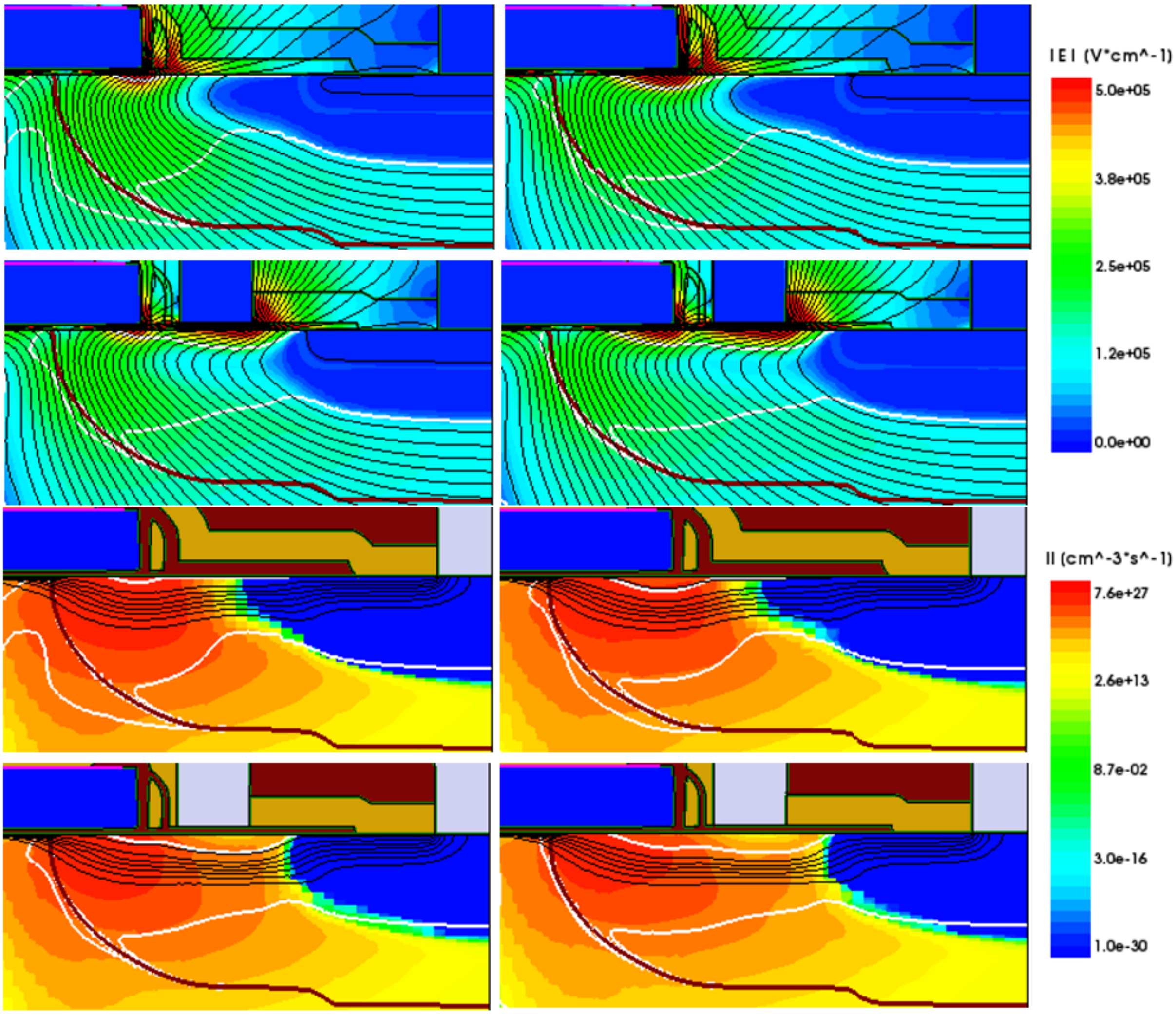

Figure 48: On-state simulations at worst-case stress conditions (at V_{GS}=2.8V and $V_{DS} \approx 12V$) of the FA- and CFP device with L_{acc}=0.2 µm (left) and L_{acc}=0.3 µm (right). The top figures show the magnitude of the electric field together with the equipotential lines. The bottom figures show the impact ionization (II) together with the current potential lines. For both devices, it is evident that increasing L_{acc} results in more diffusion into the bulk which is beneficial for HCI. Moreover, it can be seen that adding the CFP lowers the electric field at the gate edge (and thereby the impact ionization) and bends the current route further into the bulk, which both are beneficial for HCI.

From previous studies, it was determined that the minimal L_{acc} needed for sufficient reliability is 0.3 µm for the FA device. This is of course quite some length that is not effectively contributing to the BV of the device (Figure 43). Next to the potential to lower the R_{ON}A-BV trade-off, the main feature of adding the CFP to typical LDMOS devices was stated to be an improvement in the reliability in terms of HCI [8]. This is because the CFP lowers the electric field at the gate edge (and thereby the impact ionization) and bends the current route further into the bulk, which both are beneficial for HCI (Figure 48) [8]. To put this to the test, it is investigated how the CFP affects the reliability and in particular if it can be used to shorten L_{acc}.

The reliability is studied from measurements of the R_{ON} and V_T degradation as a function of time under the worst-case stress condition, as was explained in subsection 4.2.2. Both the FA- and CFP devices with varying L_{acc} (0.1, 0.2 and 0.3 µm) were measured. The worst-case stress condition was found to be at V_{DS}=12V and V_{GS}=2.8V. In Figure 49, the measurements are shown:

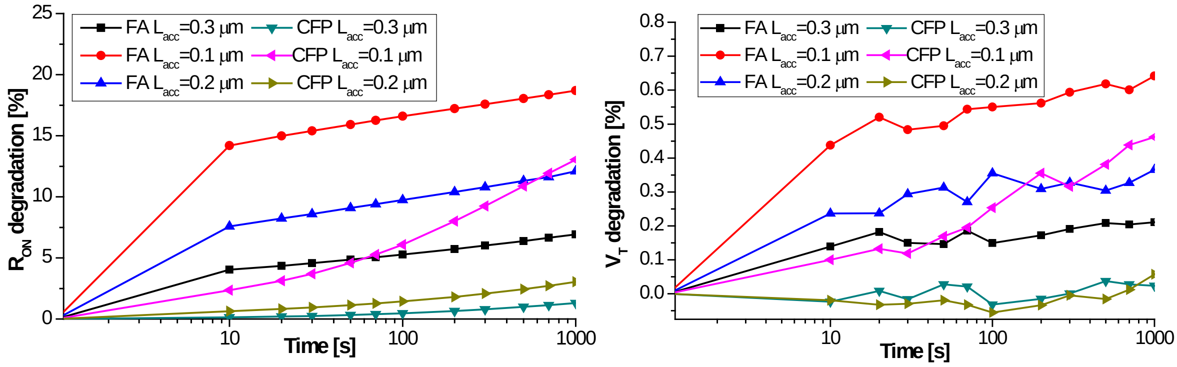

Figure 49: Reliability measurements for both the FA- and CFP devices with varying L_{acc} (0.1, 0.2 and 0.3 µm). Left: R_{ON} degradation as a function of time. Right: V_T degradation as function of time. From both graphs, it is evident that the CFP device with L_{acc}= 0.2 µm outperforms the FA device with L_{acc}= 0.3 µm. So, in terms of reliability adding the CFP gives the possibility to reduce L_{acc} to 0.2 µm.

From the R_{ON} and V_T degradation graphs in Figure 49, basically the same trends can be detected. First of all, it can be seen that the shorter L_{acc}, the larger the degradation (as expected). Secondly, adding the CFP indeed decreases the degradation (after 1000 s) by approximately 5-9% for R_{ON} and 0.2-0.3% for V_T. Lastly, the most important observation is that the CFP device with L_{acc}=0.2 µm outperforms the FA device with L_{acc}=0.3 µm. So, based on reliability, the new standard for L_{acc} can be 0.2 µm instead of 0.3 µm for the CFP device. This gives the possibility for a decrease in R_{ON}A without compromising the onset of leakage (Figure 43).

Another important aspect which may be affected by adding the CFP to the FA device, are the (dynamic and static) power losses. These will be discussed in the next subsection.

6.2.5 Power losses

When the operating frequency of a device approaches the MHz-range, charging and discharging of the device capacitances can severely degrade the dynamic (switching) power losses. In order to assess the effect of the CFP on the dynamic power losses during switching, measurements are performed of the total gate capacitance (C_{gg}), the gate-to-drain capacitance (C_{gd}) and the source-to-drain capacitance (C_{sd}), as was explained in subsection 4.2.3. In total three devices are measured: the default FA- and CFP device and the CFP device with L_{acc}=0.2 µm. In the figure below, these measurements are shown:

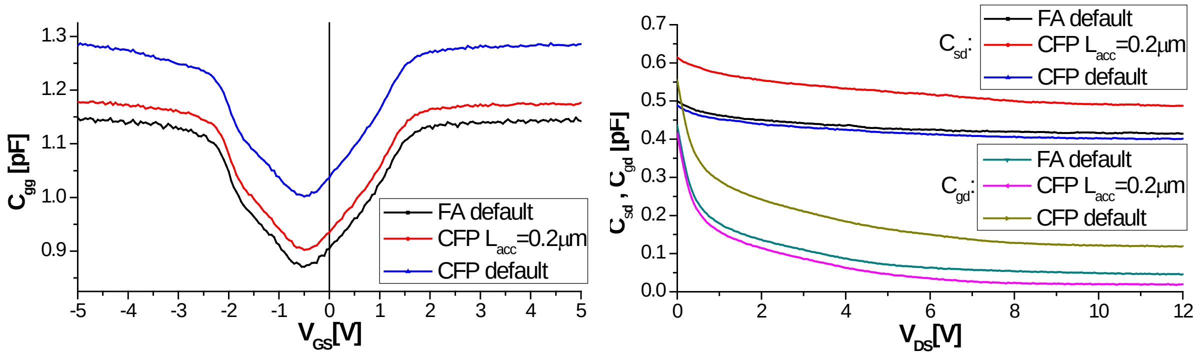

Figure 50: Capacitance measurements of the default FA- and CFP device and the CFP device with L_{acc}=0.2 µm. Left: The total gate capacitance (C_{gg}) as a function of V_{GS}. Right: the gate-to-drain capacitance (C_{gd}) and the source-to-drain capacitance (C_{sd}) as a function of V_{DS}. The measurements are performed with the following settings: 100 mV and 10 kHz for C_{gg}, 50 mV and 50 kHz for C_{gd} and 250 mV and 50 kHz for C_{sd}, all with a resolution of 50 mV.

From Figure 50 (left), it stands out that the C_{gg} curves deviate from that of a symmetric MOSFET (Figure 3 (right)). This is because additional capacitances associated with the gate overlapping a part of the n-type drain-extended region results in a shift and an increase in the overall C_{gg} curve [87]. From Figure 50 (right), it can be seen that the C_{sd} and C_{gd} curves both start at their maximum at V_{DS}=0V and converge to their minimum for $V_{DS} \rightarrow$12V as the device gets fully depleted.

In order to quantify the results from the curves in Figure 50, the charges needed for switching are extracted. These are determined by integrating from 0V (off-state) to V_{GS}=5V or V_{DS}=12V (on-state). These numbers can subsequently be used to express some commonly used FOM's. In the table below, an overview is given of the R_{ON}, the extracted charges and some important composite FOM's for these devices:

Table 7: Overview of extracted parameters and some important FOM's for the default FA- and CFP device and the CFP device with L_{acc}=0.2 µm. The R_{ON}'s are determined by $I_D - V_{GS}$ measurements and the charges (Q) are obtained by integrating the capacitance measurements. The width of these devices is 320 µm.

	FA default	**CFP default**	**CFP L_{acc}=0.2 µm**	*Unit*
R_{ON}	8.88	8.28	7.80	Ω
Q_{gg}	5.48	6.18	5.64	pC
Q_{sd}	5.18	5.04	6.27	pC
Q_{gd}	1.03	2.09	0.71	pC
$Q_{sd}+Q_{gd}$	6.21	7.13	6.98	pC
$Q_{gg}R_{ON}$	48.68	51.16	43.99	pC·Ω
$(Q_{sd}+Q_{gd})R_{ON}$	55.16	59.04	54.46	pC·Ω

To better understand the results in Table 7 (and Figure 50), schematic illustrations are shown in Figure 51 of LDMOS devices with and without a FP (attached to the gate) in which the relevant capacitances are indicated. Note that just as in the measurements, the source is shorted to the body here such that $C_{gg}=C_{gs}+C_{gd}$.

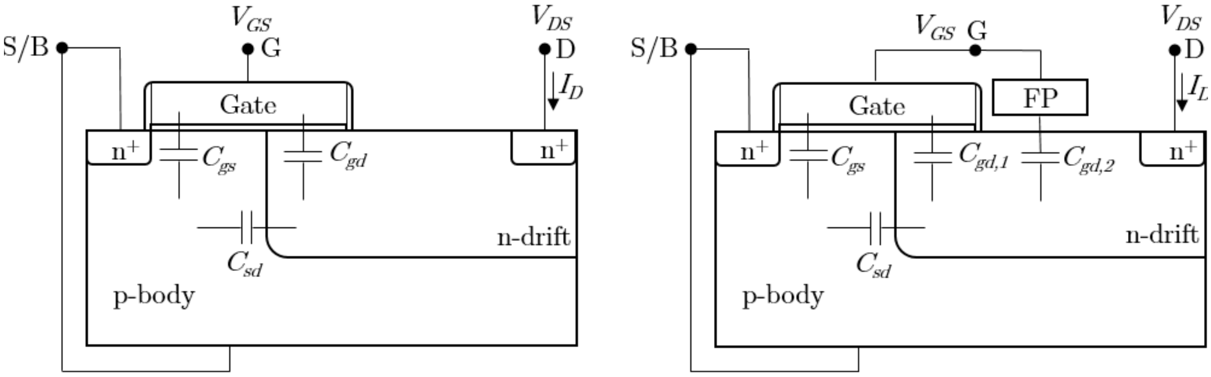

Figure 51: Schematic illustration of the relevant capacitances in LDMOS devices without (left) and with a FP attached to the gate (right). Note that the source is shorted to the body. Adding a FP (attached to the gate) results in an additional parallel gate-to-drain capacitance, such that $C_{gd}=C_{gd,1}+C_{gd,2}$.

Depending on the amount of charge needed for switching, switching the voltage of a terminal requires a certain amount of time. During this time, the device is neither in the on- or off-state, such that the power is wasted. Switching power losses are generally divided in input- and output switching losses. [1], [14]

Input switching power losses are the losses associated with charging the gate to the supply voltage and also discharging it to 0V. This is directly related to Q_{gg}, which should therefore be as small as possible to minimize input switching losses during an operation cycle. From Table 7, adding the CFP results in an increase in this charge (+12.7%). This is because the CFP results in an additional parallel gate-to-drain capacitance, such that C_{gd}=$C_{gd,1}$+$C_{gd,2}$ (Figure 51 (right). Shortening the L_{acc} subsequently reduces this to +2.9%, since this results in a decrease of $C_{gd,1}$ due to decreased overlap of the gate on the drain-extension.

Output switching power losses are the switching losses associated with voltage transitions at the drain. These transitions couple to the source and gate terminals via the C_{sd} and C_{gd} capacitances respectively, resulting in charging or discharging. The amount of charges needed for switching at these terminals (Q_{sd}+Q_{gd}) then determine the output switching power losses. From Table 7, adding the CFP results in an increase in this charge (+14.8%), which is mainly the result of the increase in Q_{gd} (due to $C_{gd,2}$). Shortening the L_{acc} only reduces this to +12.4%, which is because the decrease in Q_{gd} (due to smaller $C_{gd,1}$) is almost entirely countered by the increase in Q_{sd}. Initially, this was not expected since decreasing L_{acc} effectively decreases the junction area between the body and drain-extension thereby expected to decrease C_{sd} and Q_{sd}. Although on its itself this is not wrong, there is probably more to it related to the CFP. A possible explanation may be found in the off-state simulation of the (default) CFP device at V_{DS} ≈12V in which the hole density is indicated:

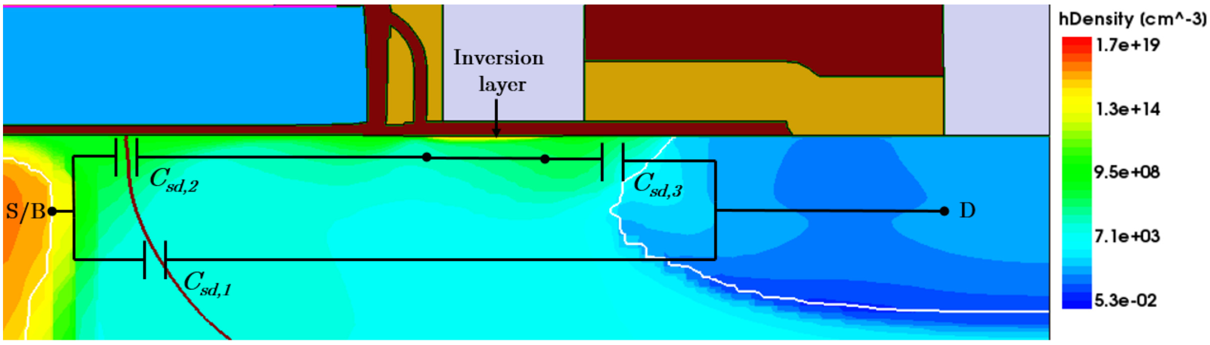

Figure 52: Off-state simulation of the default CFP device at V_{DS} ≈ 12V in which the hole density is shown. Besides the capacitance ($C_{sd,1}$) that couples directly from the drain to the body/source, due to the presence of an inversion layer below the CFP there are also capacitances ($C_{sd,2}$ and $C_{sd,3}$) that allow a coupling from the drain via the inversion layer to the body/source.

From this figure, it can clearly be seen that an inversion layer exists in the drift region below the CFP. The reason that the inversion layers forms below the CFP (and not (yet) below the gate), can be made clear by a simple hands-on calculation of the V_T of the CFP and the gate on the drain-extension. Using the version of Equation (2) suited for a n-type substrate [11] together with N_d=10^{17} cm^{-3}, $t_{ox,gate}$=13 nm, $t_{ox,CFP}$ =21 nm, $\phi_{ms,gate}$=-0.15 eV (Appendix A.1), $\phi_{ms,CFP}$ =0.39 eV [88] and T=300 K, gives: $V_{T,gate}$=-0.97V and $V_{T,CFP}$=-0.42V. The lower V_T of the CFP (which is mainly because the work function of tungsten is approximately 0.54 eV larger than that of n$^+$-polysilicon [51], [88]) in combination with the fact that the CFP is closer to the drain, results in that the inversion layer forms earlier for increasing positive drain bias below the CFP than below the gate. So as a result, besides the capacitance ($C_{sd,1}$) that couples directly from the drain to the body/source, there are now also capacitances ($C_{sd,2}$ and $C_{sd,3}$) that allow a coupling from the drain via the inversion layer to the body/source. The C_{sd} capacitance (as shown in Figure 51) is then the total circuit capacitance of the network as depicted in Figure 52, which can be expressed as: $C_{sd} = C_{sd,1} + (\frac{C_{sd,2}C_{sd,3}}{C_{sd,2}+C_{sd,3}})$. When now L_{acc} is reduced, $C_{sd,2}$ is expected to increase due to the reduced distance between the inversion layer and the body. This could therefore be the cause of the increase in C_{sd} (and Q_{sd}) for decreasing L_{acc}.

Clearly, adding the CFP is not beneficial for both types of switching losses, not even with the shorter L_{acc} that is possible for the CFP device (Table 7). However, next to the switching power losses, there is also on-state power dissipation which is proportional to the R_{ON} [1], [14]. Important FOM's which are therefore used to express the total power loss, although still separated by input- and output switching losses, are Q_{gg}R$_{ON}$ and (Q_{sd}+Q_{gd})R$_{ON}$. From Table 7, it can be seen that the Q_{gg}R$_{ON}$ and (Q_{sd}+Q_{gd})R$_{ON}$

are reduced by 9.6% and 1.3% respectively by adding the CFP and reducing L_{acc} to 0.2 µm. So, the increase in switching losses is made up by the decrease in on-state power dissipation. Nevertheless, it should be mentioned that since the switching power losses are proportional to the frequency of operation [1], [14], the CFP device performance may be impaired with respect to the FA device for high frequency operation. In case this turns out to be true and high frequency operation is required, the main parameter to minimize is the total C_{gd} (also called the Miller capacitance) [1], [14]. An easy way to reduce the C_{gd}, although at the expense of C_{ds} and C_{gs}, is by connecting the CFP to the source instead of to the gate such that the effective overlap area of the gate on the drain-extension is reduced [1]. The trade-off of this is then that the benefit of R_{ON} reduction due to additional accumulation is lost.

Now that the experimental results are discussed and valuable information is obtained regarding the influence of the CFP and the dimensions on the performance, it will be shown how the CFP devices can be optimized.

6.3 Optimization

From subsections 6.2.1 - 6.2.3 it became clear that the default CFP device, initially meant for 12V application, is far from optimized. In this section, it will first be shown by simulation how this device can be optimized. Thereafter, it will be shown by simulation that the CFP approach also can be used to effectively scale to higher voltage applications. A constraint for these optimizations is that no additional masks should be used to limit the costs. Amongst other things, this means that the doping implants and doses are limited to those available in NXP's technology platform.

6.3.1 12V application

The approach for optimizing for the initial 12V drain application, is basically improving the R_{ON}A-BV trade-off as much as possible while still ensuring that it can handle 12V at the drain at full gate drive (5V). In the figure below, it is shown how this is done by all intermediate steps on a R_{ON}A-BV graph for the default FA device. In this graph, also the measured R_{ON}A-BV points of the default FA- and CFP device are indicated, as well as the state-of-the-art fit defined in subsection 2.3.3. It should be mentioned that here the BV is defined again as the point where the $I_D - V_{DS}$ slope become infinitely.

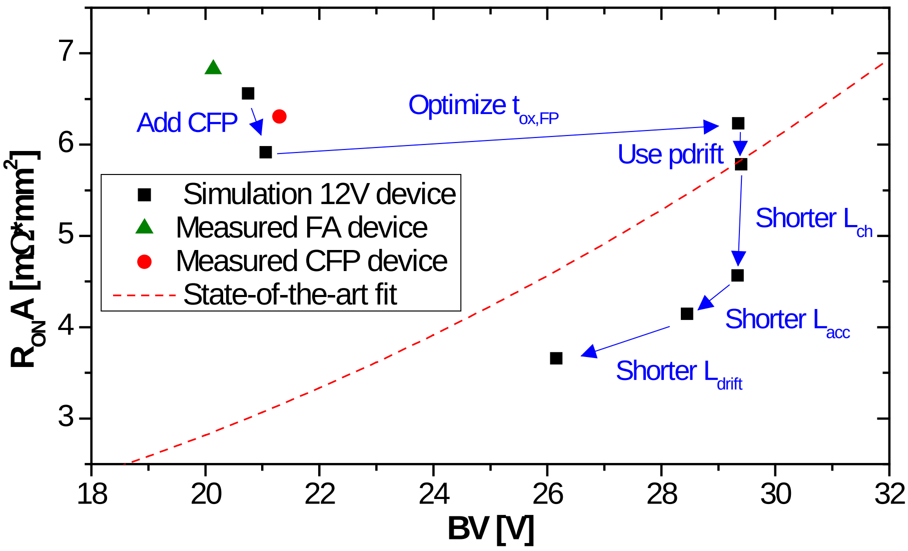

Figure 53: Optimization sequence by simulation of the 12V FA device without using any additional masks. Also, the measured R_{ON}A-BV points of the default FA- and CFP device are indicated as well as the state-of-the-art fit. The final optimized 12V device is approximately 20% below the state-of-the-art fit.

From Figure 53, it can be seen that there is a significant optimization possible without using any additional masks. Adding the CFP, results in a decrease in R_{ON}A and a slight increase in BV as already was shown in subsection 6.2.1. However, with the default $t_{ox,FP}$ setting (21 nm), the electric field distribution is far from optimal (Figure 46). By increasing $t_{ox,FP}$ to 60 nm, the BV can be increased by ~8V according to simulations. Note that this optimal $t_{ox,FP}$ for the BV is slightly bigger than the optimal for the onset of leakage (50 nm, Figure 46 (left)). Another optimization that can be done, without affecting the lateral dimensions, is using another channel implant. The channel implant that is used now (pwell) is originally optimized for GO2 devices, whereas the pdrift implant is the one optimized for GO3 devices. The pdrift implant has a lower dose than the pwell implant, such that switching to this implant reduces the V_T and Coulomb scattering and hence the R_{ch} (Equation (7)). For further minimizing the R_{ON}A, it became clear from the previous section that both the L_{ch} and L_{acc} can be reduced to 0.20 µm without sacrificing the onset of leakage (Figure 43 (right) and Figure 44 (right)). In Figure 53 however, it can be seen that reducing L_{acc} does slightly reduces the BV. This is because reducing L_{acc} reduces the distance between the field peaks at the gate and CFP edge, thereby increasing the effective distance with high electric field over which impact ionization/avalanching can occur. Lastly, since a BV of 28V for a 12V device is unnecessary high, L_{drift} can be reduced from 0.4 µm to the minimal 0.32 µm while still having sufficient BV. This puts the optimized 12V device, with a reduced pitch of 2.07 µm, approximately 20% below the state-of-the-art fit. In the figure below, the simulated $I_D - V_{GS}$ and off-state breakdown curves are shown for this device as well as for the default CFP device for reference. The cross-sections belonging to these simulations are shown in Figure 73 and Figure 74 from Appendix C.

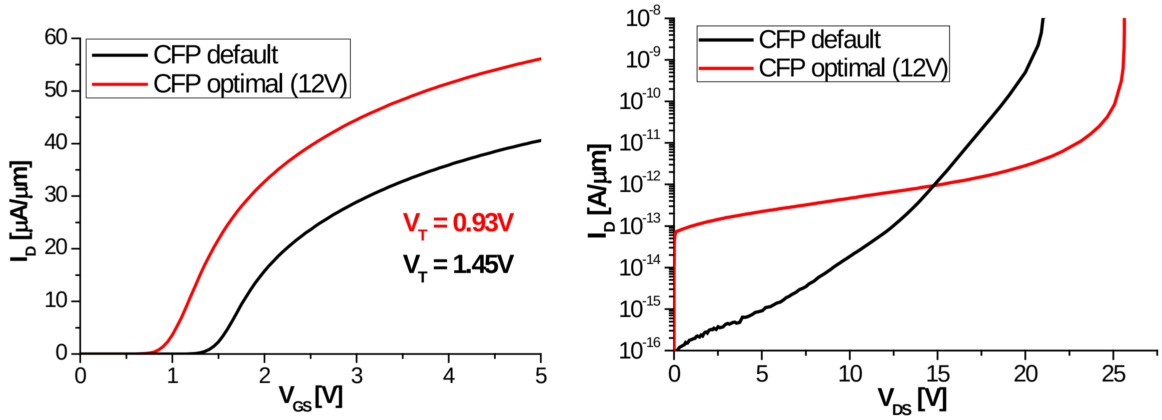

Figure 54: Simulated $I_D - V_{GS}$ (left) and off-state breakdown (right) curves for the optimized 12V CFP device (pitch=2.07 µm) as well as for the default CFP device (pitch=2.40 µm) for reference. The $I_{D,lin}$ is increased by ~38% (decrease in $R_{ON}A$ also ~38%) and the BV is increased by ~24%.

From Figure 54 (left), it can be seen that $I_{D,lin}$ is increased by ~38% (decrease in R_{ON}A also ~38%) for the optimized CFP device. Moreover, it is evident that the V_T is reduced from approximately 1.45V to 0.93V as a result of the other channel implant. From Figure 54 (right), it can be seen that the BV (at I_D=1x10^{-8} A/µm) is increased by ~24% for the optimized CFP device. However, in the low-voltage regime, it can be observed that leakage current is higher for the optimized device. This is a consequence of increased sub-threshold leakage (which scales with $\propto e^{\frac{-eV_T}{k_BT}}$ [51]) due to the lower V_T.

So, from Figure 53 and Figure 54 it is evident that both the R_{ON}A and BV can be optimized, resulting in a large improvement in the R_{ON}A-BV trade-off. However, a drawback of this optimization is the decrease in on-state breakdown (BV_{ON}), as can be seen in Figure 55. In this figure, the simulated on-state breakdown curves are shown for the default- and optimized CFP device at a V_{GS} of 2, 3.5 and 5V. The reason for lower BV_{ON} is mainly a combination of the lower R_{ON} (resulting in higher current density) and the reduced lateral dimensions (resulting in higher electric fields), which result in more impact ionization. In

Figure 75 (Appendix C), this is illustrated by simulation cross-sections of the default- and optimized CFP device at V_{GS}=5V and V_{DS} ≈12V.

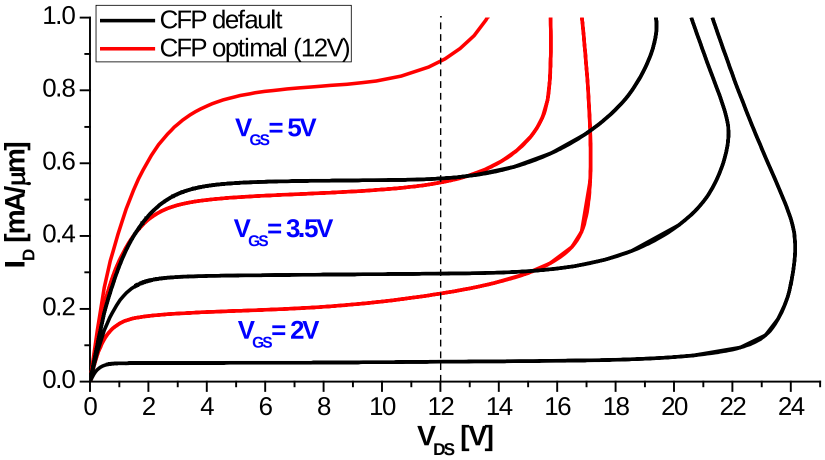

Figure 55: On-state breakdown curves for the default- and optimized CFP device at a V_{GS} of 2V, 3.5V and 5 V. The optimized device breaks down much earlier than the default device due to the lower R_{ON} and the reduced lateral dimensions, resulting in more impact ionization. Note that self-heating is not included.

The degraded on-state breakdown characteristics for the optimized device however, are not necessarily a problem. This is because the device is intended for switching applications, in which it generally operates in either the linear regime (high V_{GS}, low V_{DS}) or the blocking state (low V_{GS}, high V_{DS}) [9]. One thing that could be a problem although, would be degraded device reliability (increased HCI) due to the increased impact ionization in the on-state. Lastly, to complete the discussion of the optimized 12V device, it is expected that the optimization is beneficial for the power losses. The decrease in R_{ON} reduces the static power losses and the expected decrease in the capacitances reduce the switching losses. The C_{gg} and C_{gd} are expected to decrease due to the large decrease in gate area and the C_{sd} is expected to decrease due to the increased $t_{ox,FP}$ (thereby reducing the inversion below the CFP (Figure 52)) .

Now that it is shown how the CFP device can be optimized for the initial 12V application to beat the current state-of-the-art, it will be shown that the CFP concept also can be used to scale effectively to higher (off-state breakdown) voltages.

6.3.2 Higher voltage applications

Generally, scaling of LDMOS devices to higher (off-state breakdown) voltages can be done by increasing the L_{drift} dimension in combination with appropriate decrease in drift doping to enable full depletion. Due to the limitation of the use of existing doping implants, this cannot be done for the CFP devices. However, because of the additional vertical depletion due to the CFP, the maximum L_{drift} for which full depletion occurs is extended with respect to the FA devices (Figure 42). In Figure 56, the effect of solely increasing L_{drift} on the BV, R_{ON}A and potential through the device at off-state breakdown (I_D=1x10^{-8} A/µm) is shown. The CFP device here has the optimized settings found for the optimal 12V device, except for L_{acc} that is restored to 0.3 µm to have sufficient reliability for higher voltages (subsection 6.2.4). From Figure 56 (left), it can be seen that the BV effectively increases until ~0.6 µm and thereafter saturates. This saturation occurs because the condition at which breakdown occurs changes from PT to NPT (subsection 2.2.2). In other words, the depletion width at which breakdown occurs ($W_{D.bd}$) becomes smaller than L_{drift}, which is evident from Figure 56 (right). Increasing L_{drift} from 0.4 to 0.6 µm increases $W_{D.bd}$ also by ~0.2 µm such that the allowable voltage drop increases approximately linear with L_{drift}. From 0.6 to 0.8 µm, the increase in $W_{D.bd}$ (and thus in the allowable voltage drop) already becomes smaller and completely vanishes for 0.8 to 1.0 µm.

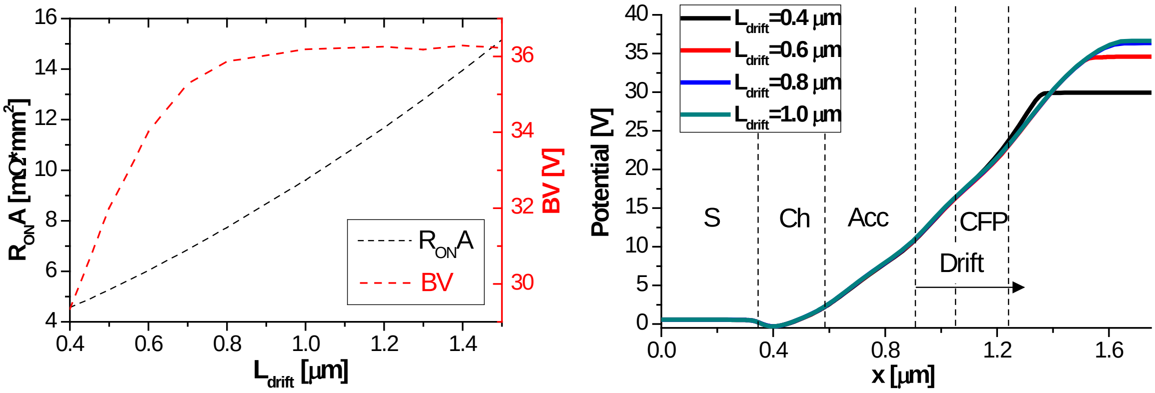

Figure 56: Effect of increasing L_{drift} while everything else fixed. The settings used here are the same as the optimized 12V device except for that L_{acc}= 0.3 µm for reliability purposes. Left: $R_{ON}A$ and BV as a function of L_{drift}. Right: Potential as a function of the lateral position through the device at off-state breakdown (I_D=1x10^{-8} A/µm). The BV effectively increases until~0.6 µm and thereafter saturates. This saturation occurs because the device breaks down before the drift region is fully depleted to the drain.

So, with the current CFP design, the maximum BV that can be reached is approximately 36V, while the goal is to scale effectively towards ~50V. Generally, the way to go, without changing the doping implants, would then be scaling of the CFP length with increasing L_{drift}. In Figure 57 (left), it is shown that this indeed can be used to get to higher BV's. The CFP length is scaled here accordingly to the drift region for L_{drift} >0.6 µm. Moreover, also the optimal $t_{ox,FP}$ increases for increasing L_{drift}. The reason for this is that for increasing CFP length, the RESURF efficiency becomes worse. In subsection 2.3.1, it was discussed that ideal RESURF using a FP only can be obtained by grading one of the parameters (e.g. $N_d(x)$ or $t_{ox,FP}(x)$, Equation (22)). Since this is not done for the CFP devices, this non-ideality becomes more and more pronounced for increasing CFP length. Increasing $t_{ox,FP}$ then slightly counters this effect by decreasing the influence of the CFP, however is still not sufficient to retain the steep BV-L_{drift} slope observed for low L_{drift}. This poorer RESURF efficiency for increasing CFP length is made evident in Figure 57 (right), where the potential and lateral electric field are plotted through the devices with L_{drift}= 0.7 µm and 1.5 µm. From this figure, it can be seen that the relative voltage drop over the CFP becomes less for the larger CFP length (1.1 µm) than for the smaller CFP length (0.3 µm). This is a consequence of the two field peaks being further apart resulting in a larger dip in-between them.

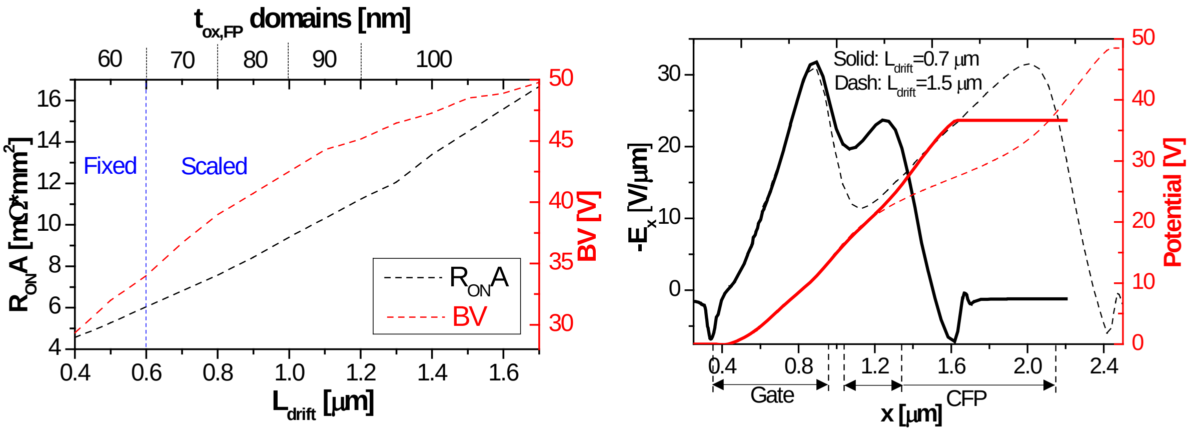

Figure 57: Left: Scaling of the CFP device up to BV's of~50V by scaling the CFP length according to the drift region for L_{drift} > 0.6 µm. In addition, also the optimal $t_{ox,FP}$ increases for increasing L_{drift}. Right: Lateral electric field and potential through the devices with L_{drift} =0.7 µm (CFP length = 0.3 µm) and 1.5 µm (CFP length = 1.1 µm). The relative voltage drop over the CFP becomes less for the larger CFP length because the two field peaks are further apart, indicating poorer RESURF efficiency.

One important thing that should be mentioned however, is that the mask used for contacts is fixed in the x-direction (lateral) and only can be scaled in the y-direction (width of the device). So, the scaling of the CFP as is, would require a dedicated mask for the FP. A simple but novel way to bypass this, is by rotating the mask and placing an array of CFP slots in the width of the device. In the figure below, this concept is illustrated using a top view of the CFP device. Although this slightly changes the initial CFP concept, since the FP is not continuous anymore in the y-direction, it does provide a fully arbitrability scalable FP.

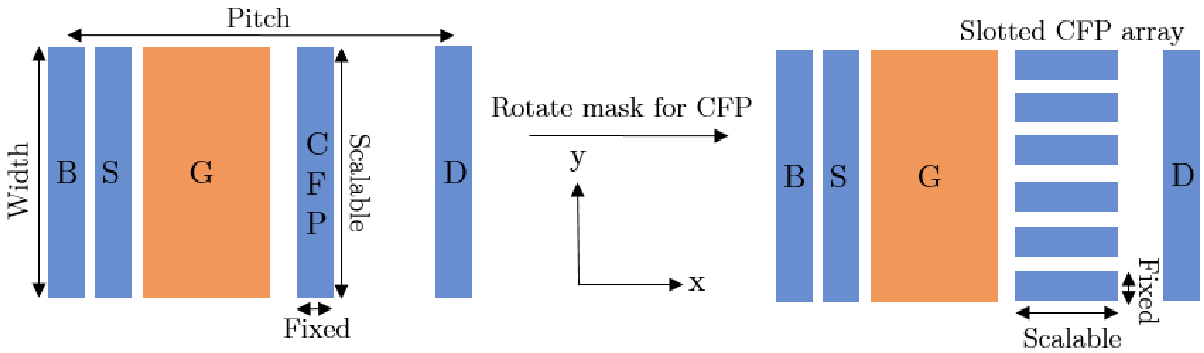

Figure 58: Novel way of scaling to higher voltages using the CFP approach without using any additional masks. Left: Top view of the CFP device in which the lateral dimension (x) of the CFP is fixed. Right: Top view of the CFP device after the mask used for the CFP is rotated 90° and used to form a slotted CFP array which is scalable in the lateral dimension.

In order to check whether the slotted CFP array indeed can be used to scale (effectively) to higher BV's, 3D simulations are set up of the scaled device with L_{drift}=1.5 µm (BV=48.5V in 2D (Figure 57 (left))). However, because 3D simulations are computationally very intensive and generally suffer from poor convergence, especially for processed devices, a simplified version of the device is mimicked using the Sentaurus™ Structure Editor tool [89]. In this version, analytical doping profiles are used and materials other than the necessary silicon and oxide are left out of the simulation. In the figure below, this structure (with merged source-body construction) is shown for three different FP constructions (left: no FP, middle: full FP and right: partial FP):

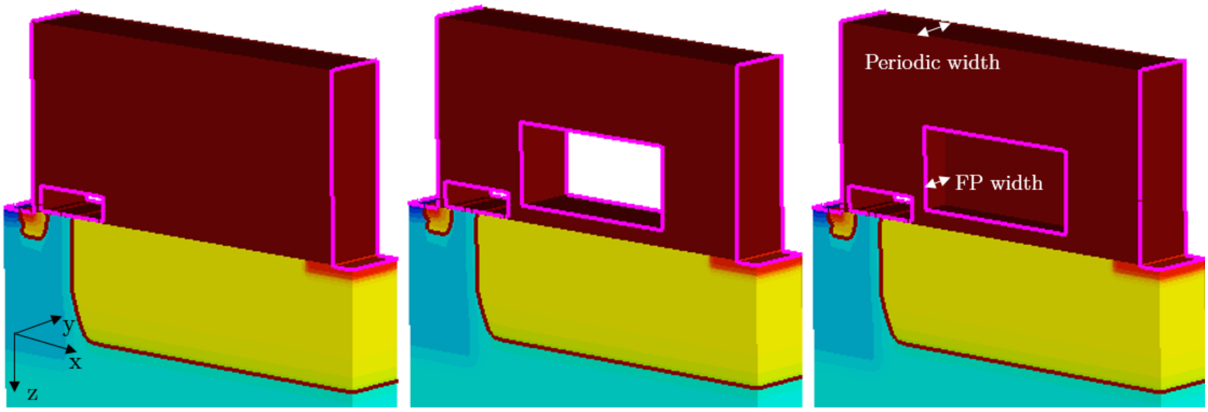

Figure 59: Simulated 3D devices (with merged source-body construction) using the Sentaurus™ Structure Editor tool with different three FP constructions (left: no FP, middle: full FP and right: partial FP). Here the color in the silicon indicates the doping concentration (red: n-type, blue: p-type). The important dimensions are: L_{ch}=0.20 µm, L_{acc}=0.30 µm, L_{drift}=1.50 µm and FP length=1.10 µm.

Device simulations that are executed with these 3D structures are done with periodic boundary conditions applied to the front- and backside. In these simulations, the FP width to periodic width ratio ($\equiv \tau$) is varied to give clear insight on how and to what extent the partial FP influences the R_{ON}A and BV.

First, off-state breakdown simulations are performed for a range of $t_{ox,FP}$ to extract the optimal $t_{ox,FP}$ (for off-state breakdown) as a function of τ. In Figure 60, the result of this is shown as well as the corresponding relative BV. In addition, also the relative BV is indicated for the structure without FP.

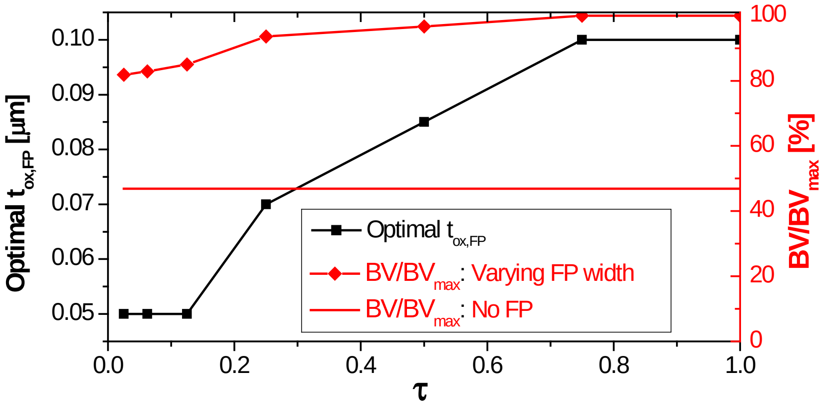

Figure 60: Optimum $t_{ox,FP}$ (for off-state breakdown) and corresponding relative BV for varying τ($\equiv$ FP width/periodic width). Also, the relative BV is indicated for the structure without FP. The optimum $t_{ox,FP}$ decreases for decreasing τ to (partly) compensate for the decreasing influence of the FP and the corresponding BV decreases towards ~80% of its original value for the full FP (τ=1). Nevertheless, it is evident that the partial FP can be used to scale to higher voltages

From this figure, it is evident that the optimal $t_{ox,FP}$ of 0.1 µm for τ = ~ 0.8 - 1 decreases to 0.05 µm for τ = ~ 1/40 - 1/8. The reason for this is that as τ becomes smaller, the relative influence of the FP becomes smaller, which can be countered by reducing $t_{ox,FP}$. However, as can be seen, this is only good for a partial compensation because for decreasing τ the corresponding BV decreases towards ~80% of its original value for the full FP (τ=1). The reason for the decreasing BV with decreasing τ is made visible in Figure 61, where cross-sections are shown (at V_{DS} ≈30V) for τ=1/8, 1/2 and 1. The cross-sections are taken in the x,y-plane through the drain-extension (left) and in the y,z-plane through the FP near the drain (right).

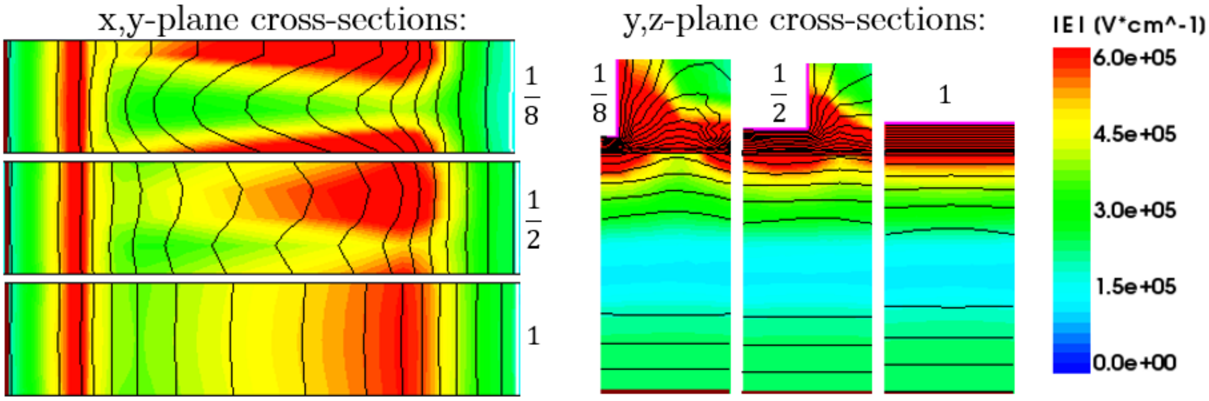

Figure 61: Cross-sections of the 3D off-state breakdown simulations at V_{DS} ≈30V for τ=1/8, 1/2 and 1. The colors indicate the magnitude of the electric field and the black lines the equipotential lines. Left: Cross-sections of the drain-extension in the x,y-plane. Right: Cross-sections through the FP (near the drain) in the y,z-plane. For decreasing τ, the volume with high field increases resulting in enhanced impact ionization and hence lower BV.

From these, it is evident that by going to a partial FP the field not only peaks at the FP edge near the drain, but also has a high field tail below the partial FP resulting in a droplet-like shaped electric field distribution in the x,y-plane. This is the result of that the partial FP has additional edges in the y-direction where field peaks arise. Moreover, as τ becomes smaller, the field peaks at both edges merge more together resulting in a stronger and more concentrated field with a longer tail. So, for decreasing τ, the volume with high electric field increases resulting in enhanced impact ionization and hence lower BV. Nevertheless, even for small τ, it is evident from Figure 60 that the partial FP can be used to scale to higher voltages,

Secondly, $I_D - V_{GS}$ simulations are performed to extract the corresponding relative R_{ON}A dependence and is shown in Figure 62. Here also the dependencies are shown when $t_{ox,FP}$ is kept constant at 0.10 µm and without FP. From this figure, it is evident that as τ becomes smaller, $R_{ON}A$ only increases up till ~3% (for optimal $t_{ox,FP}$) due to decreased accumulation. This small increase is on one hand due to the thinner (optimal) $t_{ox,FP}$ for smaller τ and on the other hand it can be seen that even when $t_{ox,FP}$ is kept constant $R_{ON}A$ only increases up to ~6%. So, even for small τ, a significant improvement in $R_{ON}A$ can be accomplished by the partial FP due to accumulation.

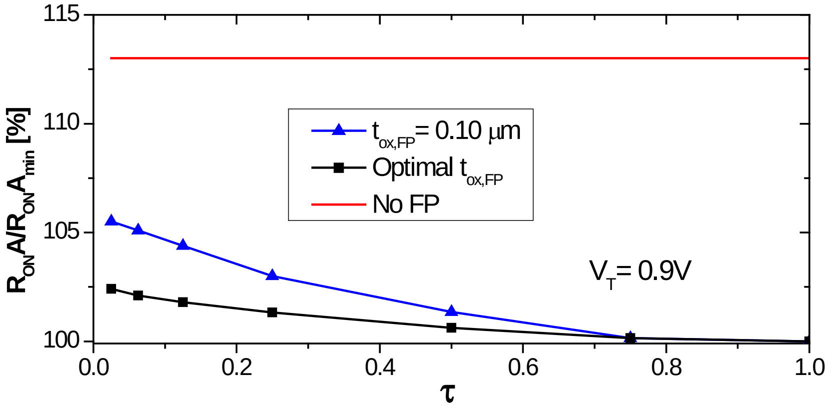

Figure 62: Relative $R_{ON}A$ as a function of τ for optimal $t_{ox,FP}$, $t_{ox,FP}$ constant at 0.10 µm and no FP. For decreasing τ, only small increases in $R_{ON}A$ are observed for the partial FP, especially when $t_{ox,FP}$ is reduced with decreasing τ (optimal $t_{ox,FP}$). So, a significant improvement in $R_{ON}A$ can be accomplished by the partial FP due to accumulation.

So, from Figure 60 and Figure 62, it can be concluded that for the most optimal performance τ should be as close as possible to 1. The maximal τ for the CFP slot array depends on how far the CFP slots minimally should be apart, which by rule of design is 0.288 µm. Moreover, the contact width by layout is 0.16 µm such that the periodic width is 0.448 µm. The τ then depends on the actual CFP width, which is approximately ~0.2 µm (Table 4), such that $\tau \approx 0.2/0.448 = 0.45$. From Figure 60 and Figure 62, this then means that the optimal $t_{ox,FP}$ is approximately 0.08 µm and that the corresponding BV and $R_{ON}A$ are decreased by ~4% and increased by ~1% respectively with respect to the full FP. Lastly, it should be mentioned that the proposed slotted CFP array architecture also provides a way to control the optimal $t_{ox,FP}$ by varying τ in the layout. This therefore allows for producing (optimized) devices targeted for different voltage applications using the same $t_{ox,FP}$ and therefore the same processing (i.e. on a single wafer).

Now that it is shown that the CFP concept effectively can be used for optimization for the initial 12V application, as well as for higher (off-state breakdown) voltage applications, an overview of the optimized devices will be shown and compared to industrial devices.

6.3.3 Overview - R$_{ON}$A-BV

In the figure below, it is shown how the optimized CFP devices compare to the state-of-the-art fit defined in subsection 2.3.3. Here the scaled devices (BV's > 35V) are corrected for the partial FP (τ=0.45). Moreover, for reference, also the R$_{ON}$A-BV data from the devices in NXP's technology platform as well as the best performing CFP devices on this wafer lot (from measurements and simulations) are indicated. The best performing CFP devices on this wafer lot have the thickest $t_{ox,FP}$ variation (~38 nm) and correspond to (from left to right) the variations with L_{drift}=0.32 µm, L_{acc}=0.20 µm, L_{drift}=0.40 µm and L_{drift}=0.50 µm.

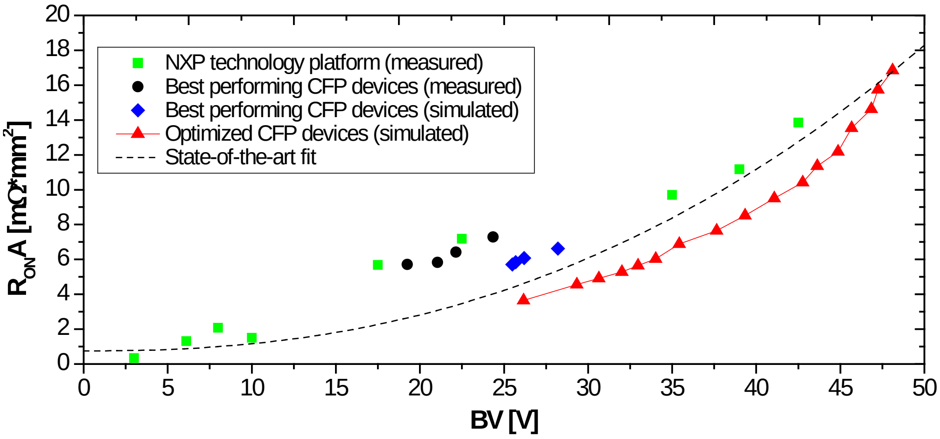

Figure 63: R$_{ON}$A-BV data of the: devices in NXP's technology platform, best performing CFP devices in this wafer lot (measured and simulated) and optimized CFP devices. Also, the state-of-the-art fit is indicated. The CFP concept provides a way to beat the best performing devices in the industry in the range from about 25V to 48V off-state breakdown.

From the simulated optimized CFP devices, it can be seen that the CFP concept provides a way to outperform the current industry standard in the range from about 25V to 48V off-state breakdown. Although, it should be noted that there is a clear offset in BV between measurements and simulations. This may result in that the real applicable range is somewhat lower. Nevertheless, to scale effectively further downwards, L_{drift} should be further downscaled, however it is already at its minimum (0.32 µm) for the optimized 12V device. On the other hand, for BV's above this range, the non-ideal FP construction results in too poor RESURF efficiency to scale effectively to higher voltages. Therefore, outside this range other device architectures would be needed. The largest improvement with respect to the state-of-the-art fit can be observed at about 41V according to the simulation, where the R$_{ON}$A is ~28% lower. In this range however, NXP's devices are already competitive with the industry. Therefore, optimizing these devices (which are other architectures) may possibly be more beneficial, both in time/monetary investment and in R$_{ON}$A-BV trade-off, than extending the CFP concept. In the range from about 17.5V to 25V however, the NXP devices are quite inferior to the industry. For this range, according to the simulation, the optimized 12V device reduces the R$_{ON}$A by ~40% while maintaining the BV and potentially beating the state-of-the-art by ~20%. Therefore, this device offers a good solution to compete with state-of-the-art devices in that range.

Chapter 7: Conclusion

In summary, the objective of this thesis was to improve integrated power devices in the existing NXP technology platform to obtain superior $R_{ON}A$-BV trade-off with respect to state-of-the-art devices in the range up to 50V off-state breakdown. For this purpose, two novel concepts, namely oxygen-inserted layer (OIL)- and contact field plate (CFP) technology, are explored by both simulation and experiment.

For the OIL approach, a combined wafer lot was set up with original wafers and wafers with the OIL, on which the devices from NXP's technology platform were contained (section 5.1). The thermal budget and channel doping implant adjustments that were needed for the OIL integration were supported by TCAD simulations. After the fabrication of the wafer lot, from PCM data the best performing device was extracted and subsequently analyzed in more detail. This included an evaluation of the linear drain current improvement for the various wafer split variations as well as for varying dimensions of the device. Moreover, to assess whether the OIL has the expected effect on the channel mobility (subsection 2.4.2) and gate leakage (subsection 2.4.5), μ_{ch}-E_{eff} and $\ln(I_G)$- t_{ox} curves were extracted. Lastly, to assess whether the OIL is integrated correctly and has the desired effect on the channel doping (subsection 2.4.4), TEM device cross-sections and SIMS profiles through the channel (of oxygen and boron) were requested. These were subsequently compared to what was expected from simulations and based on that could be used to clarify some of the observations made in the electrical analysis. Moreover, additional supportive information was obtained by a comparison of the current experiment with those in literature [4]–[7], [70]. Nevertheless, the experiments conducted in this work concerning the OIL concept are not conclusive yet and follow-up work is therefore necessary.

For the CFP approach, a wafer lot was available in which both the FA- and CFP devices were included with varying dimensions (section 6.1). From electrical measurements and simulations (section 6.2), it became evident that the CFP devices were far from optimized. For the BV, the crucial parameter turned out to be the oxide thickness below the CFP ($t_{ox,FP}$), which in this wafer lot was believed to be significantly below the optimum. For the $R_{ON}A$, significant room for improvement is in reducing the channel length. This also became evident from separating the resistance contributions, which showed a relatively high channel contribution to the resistance of ~40%. Additionally, from reliability measurements it became clear that by adding the CFP the accumulation length (L_{acc}) can be reduced by 0.1 μm without degrading lifetime. Moreover, for the reduced L_{acc}, the increase in switching losses by adding the CFP is made up by the decrease in on-state power dissipation. Finally, to show how the CFP devices can be optimized, without using any additional masks, optimizations are proposed by simulations for the initial 12V application as well as for higher voltages (section 6.3). For the 12V application, the optimized dimensions and other channel implant reduce the $R_{ON}A$ by ~40% resulting in a device that has the potential to beat state-of-the-art devices by ~20%. Using the same architecture, BV's up to ~35V can effectively be achieved by appropriate scaling of the drift region length and $t_{ox,FP}$. Scaling effectively to higher voltages requires a novel slotted CFP array architecture, with which state-of-the-art outperforming devices can be obtained up to BV's of about ~48V.

So, simulation results showed that utilizing the CFP concept and subsequent optimization provides a way to beat current state-of-the-art integrated power devices (for BV's<50V) in the existing NXP technology platform. Moreover, if the OIL works as previously published [2]–[7], this concept has the potential to add additional improvement. In order to demonstrate this on silicon, follow-up cycles are needed for both concepts. When the proposed optimizations from this work are followed, the CFP concept alone already has the potential to beat the current state-of-the-art in the range from about ~25V to ~48V off-state breakdown.

Chapter 8: Outlook

The work reported on in this thesis basically covers one full silicon cycle, setting up a wafer lot, production by the fabric and finally the analysis and evaluation. Although the semiconductor industry is relatively mature and extensive simulation tools are available to predict the processing of a device and its characteristics, realizing it on silicon remains challenging. Generally, designing new devices or improving existing devices therefore requires multiple cycles before it can be industrialized. The two concepts that are explored in this work have just completed the first cycle, such that there is still much room for improvement. In this chapter, it will be discussed what is recommend for follow-up work, including additional silicon cycles.

For a potential follow-up cycle, it is first of all recommended to include the optimized 12V device (subsection 6.3.1). Secondly, it is recommended to include some devices targeted for higher voltage applications (subsection 6.3.2) to show that the CFP concept is also feasible for higher voltages. A proposal for a selection of devices with typical application targets (based on the simulations) are shown in the table below.

Table 8: Best guess dimensions for CFP devices for several typical target applications with BV's below 50V. Also, it is indicated whether the fixed CFP should be used or the scalable slotted CFP array. All dimensions are in μm, unless stated otherwise.

BV [V]	R_{ON}A $[m\Omega \cdot mm^2]$	CFP	CFP length	$t_{ox,FP}$	L_{ch}	L_{acc}	L_{drift}
26	3.7	Fixed	0.20	0.06	0.20	0.20	0.32
32	5.3	Fixed	0.20	0.06	0.20	0.30	0.50
39	8.4	Scaled	0.50	0.06	0.20	0.30	0.90
47	14.5	Scaled	1.10	0.08	0.20	0.30	1.50

Besides implementing these devices with these particular dimensions, also variations should be incorporated in which each of the parameters is independently varied. In that manner, clearly the influence of each parameter on the device characteristics as well as discrepancies with the simulations can be mapped. This is especially important for $t_{ox,FP}$, because small variations already can have a large effect on the BV (Figure 46). In addition, for reliability purposes it is recommended for the higher voltage devices to incorporate sufficient upward L_{acc} variations (up to ~0.60 μm), as it is not clear what is needed for sufficient reliability.

In subsection 6.3.2, it was mentioned/shown that for increasing length of the FP the RESURF efficiency becomes poorer. Which is a result of the constant $t_{ox,FP}$ instead of an ideally linearly increasing $t_{ox,FP}$ along the drain-extension (Equation (22)). A potential way to improve on this this and thus to scale more effectively to higher voltages using the CFP concept, is shown in Figure 64. Here the standard (continuous) CFP is combined with the slotted CFP array. Since the slotted CFP array basically acts as a (continuous) FP with a larger effective $t_{ox,FP}$, a "stepped" FP structure is obtained by layout rather than by processing. This stepped CFP structure more closely resembles the ideal FP structure, such that better RESURF conditions can be obtained [9], [10], [16]. Before producing such an architecture on silicon however, its feasibility and optimal dimensions should be explored by simulation first.

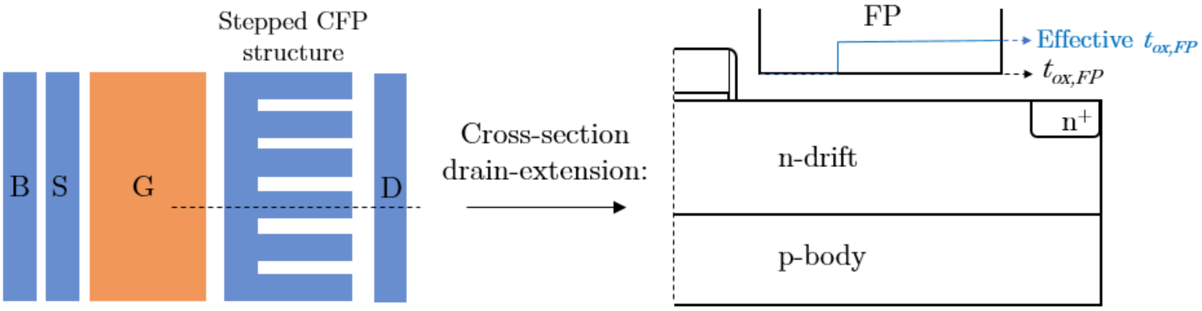

Figure 64: Proposed stepped CFP architecture to scale more effectively to higher voltages. Here the standard CFP is combined with the slotted CFP array. Since the slotted CFP array basically acts as a (continuous) FP with a larger effective $t_{ox,FP}$, a "stepped" FP is obtained allowing for better RESURF [9], [10], [16].

Besides the optimizations discussed in this work for the CFP devices, of course there is also room for additional improvements. Some possibilities that are worth investigating in a follow-up cycle (which are not inherently limited to the CFP devices) are listed below:

- Segmented source-body construction: For the devices in this work, the implants for the source and body contacts were aligned laterally (x-direction), either in contact with each other (for a merged source-body contact) or separated by a STI (for separated source and body contacts). From earlier experiments at NXP and literature (e.g. [90]), it is proven that these implants also can be segmented periodically in the width of the device (y-direction). The width of the source and body contact implants are 0.42 µm, such that by going to the segmented source-body construction the device pitch can be reduced by 0.42 µm. For the optimized 12V CFP device, whose pitch is 2.07 µm, this could lead to a ~20% reduction in R_{ON}A. Combining this technique with short channels however, could possibly be problematic (in the on-state) when the current has not enough room to spread out.
- Reducing drain length: From Figure 33, it can be seen that the drain implant takes quite some pitch, specifically 0.56 µm. The part of which is below the (SiProt) oxide layer (~0.20 µm), is to ensure that the SiProt layer covers the whole drift region also in case of large process variation. This could likely be reduced by ~0.10 µm and still provide sufficient overlap. For the optimized 12V CFP device, this could lead to a ~5% reduction in R_{ON}A.
- Self-aligned channel implant: The current process variation in the channel length can be as large 0.10 µm. For small channel lengths (such as the 0.20 µm proposed for the CFP devices), this can be problematic due to the V_T roll-off (Figure 45). Therefore, it is recommended to put effort in minimizing this process variation by going to a new channel implant. For that purpose, a self-aligned channel implant could offer a solution. The channel is then implanted under an angle below the gate where the gate itself acts as a mask, instead of implanting it through a photoresist mask before the gate formation (see for example Figure 22). This reduces misalignment with respect to the gate, thereby reducing process variation in the channel length.
- OIL concept: Although this concept is already extensively explored in this work, it cannot yet be concluded whether it is compatible with the devices in NXP's technology platform. Therefore, it still remains a potential candidate for lowering the R_{ON}A-BV trade-off. In addition, this technique could also aid the step to shorter channels, as it is claimed to improve short channel control [2]. Lastly, in case of a follow-up cycle regarding this concept, it is recommended to combine it also with the CFP devices. Since these devices have additional accumulation below the FP in the on-state, the (R_{ON}A) improvement is expected to be higher than that for the regular FA devices.

Bibliography

[1] T. Erlbacher, *Lateral Power Transistors in Integrated Circuits*, 1st ed. Springer, 2014.

[2] N. Damrongplasit *et al.*, “Comparative Study of Uniform Versus Supersteep Retrograde MOSFET Channel Doping and Implications for 6-T SRAM Yield,” *IEEE Trans. Electron Devices*, vol. 60, no. 5, pp. 1790–1793, 2013.

[3] R. J. Mears *et al.*, “Simultaneous carrier transport enhancement and variability reduction in Si MOSFETs by insertion of partial monolayers of oxygen,” *2012 IEEE Silicon Nanoelectron. Work. SNW 2012*, vol. 50, pp. 1–2, 2012.

[4] N. Xu *et al.*, “Electron mobility enhancement in (100) oxygen-inserted silicon channel,” *Appl. Phys. Lett.*, vol. 123502, no. 107, pp. 2–6, 2015.

[5] N. Xu *et al.*, “MOSFET performance and scalability enhancement by insertion of oxygen layers,” *Tech. Dig. - Int. Electron Devices Meet. IEDM*, pp. 127–130, 2012.

[6] N. Xu *et al.*, “Effectiveness of Quasi-Confinement Technology for Improving P-Channel Si and Ge MOSFET Performance,” *IEEE*, pp. 2–3, 2012.

[7] N. Xu *et al.*, “Extension of planar bulk n-channel MOSFET scaling with oxygen insertion technology,” *IEEE Trans. Electron Devices*, vol. 61, no. 9, pp. 3345–3349, 2014.

[8] L. Wei, C. Chao, U. Singh, R. Jain, L. Leng.Goh, and P. R. Verma, “A Novel Contact Field Plate Application in Drain-Extended-MOSFET Transistors,” in *Proceedings of the 29th International Symposium on Power Semiconductor Decives & ICs, Sapporo*, 2017, pp. 1–3.

[9] B. K. Boksteen, “Field-plate assisted RESURF power devices: Gradient based optimization, degradation and analysis,” University of Twente, 2015.

[10] A. Ferrara, “RESURF power semiconductor devices: performance and operating limits,” University of Twente, 2015.

[11] D. A. Neamen, *Semiconductors Physics and Devices*, 3d ed. University of New Mexico: McGraw-Hill, 2012.

[12] G. E. Moore, “Cramming more components onto integrated circuits,” *Proc. IEEE*, vol. 86, no. 1, pp. 82–85, 1965.

[13] N. Mohan, T. M. Undeland, and W. P. Robbins, *Power Electronics - Converters, Applications and Design*, 2nd ed. John Wiley & Sons, 1995.

[14] B. J. Baliga, *Fundamentals of Power Semiconductor Devices*, 1st ed., vol. 53, no. 9. North Carolina: Springer, 2008.

[15] J. A. Appels and H. M. J. Vaes, “High Voltage Thin Layer Devices (RESURF Devices),” *IEEE*, pp. 238–241, 1979.

[16] A. Tannenbaum, D. Mistele, Y. Stav, M. H. Emek, and P. O. Box, “Optimization of Integrated 0.18 um nLDMOS, for Power Management ICs Rated at 40-60V,” in *IEEE International Conference on Microwaves, Antennas, Communications and Electronic Systems*, 2017, pp. 2–6.

[17] T. Mori, H. Fujii, S. Kubo, and T. Ipposhi, “Investigation into HCl improvement by a split-recessed-gate structure in an STI-based nLDMOSFET,” in *Proceedings of the International Symposium on Power Semiconductor Devices and ICs*, 2017, pp. 459–462.

[18] S. Chung, “Analytic model for field-plate-edge breakdown of planar devices terminated with field plate and semiresistive layer,” *IEEE*, vol. 151, no. 1, pp. 21–24, 2004.

[19] L. Vestling, B. Edholm, J. Olsson, S. Tiensuu, and A. Soderbag, “a Novel High-Frequency High-Voltage Ldmos Transistor Using an Extended Gate Resurf’ Technology,” *Int. Symp. Power Semicond. Devices IC’s*, pp. 45–48, 1997.

[20] J. Wei *et al.*, “High-Voltage Thin-SOI LDMOS with Ultralow ON-Resistance and even Temperature Characteristic,” *IEEE Trans. Electron Devices*, vol. 63, no. 4, pp. 1637–1643, 2016.

[21] I. Y. Park *et al.*, “Implementation of buffered super-junction LDMOS in a 0.18um BCD process,” in *Proceedings of the International Symposium on Power Semiconductor Devices and ICs*, 2009, pp. 192–195.

[22] J. Somayaji, B. S. Kumar, M. S. Bhat, and S. Member, “Performance and Reliability Codesign for Superjunction Drain Extended MOS Devices,” *IEEE Trans. Electron Devices*, vol. 64, no. 10, pp. 1–9, 2017.

[23] J. Sonsky and A. Heringa, “Dielectric resurf: breakdown voltage control by STI layout in standard CMOS,” *IEDM Tech. Dig.*, vol. 00, no. d, pp. 373–376, 2005.

[24] A. Heringa, J. Šonský, J. Perez-Gonzalez, R. Y. Su, and P. Y. Chiang, “Innovative lateral field plates by gate fingers on STI regions in deep submicron CMOS,” in *Proceedings of the International Symposium on Power Semiconductor Devices and ICs*, 2008, pp. 271–274.

[25] R. Y. Su *et al.*, “On-state resistance improvement by partially slotted STILDMOS transistor in 0.25-micron smart power technology,” in *International Conference on Solid-State and Integrated Circuits Technology Proceedings*, 2008, pp. 199–202.

[26] P. Kundu and R. Yadav, “Effect of High-K Dielectric Materials on Leakage Current,” *Int. J. Electron. Comput. Sci. Eng.*, vol. 1, no. 3, pp. 1454–1458, 2009.

[27] J. Li, P. Li, W. Huo, G. Zhang, Y. Zhai, and X. Chen, “Analysis and fabrication of an LDMOS with high-permittivity dielectric,” *IEEE Electron Device Lett.*, vol. 32, no. 9, pp. 1266–1268, 2011.

[28] A. Ozbek, “Measurement of Impact Ionization Coefficients in Gallium Nitride,” Faculty of North Carolina State University, 2012.

[29] J. He *et al.*, “Linearly graded doping drift region: A novel lateral voltage-sustaining layer used for improvement of RESURF LDMOS transistor performances,” *Semicond. Sci. Technol.*, vol. 17, no. 7, pp. 721–728, 2002.

[30] W. Ge *et al.*, “Ultra-low on-resistance LDMOS with multi-plane electron accumulation layers,” *IEEE Electron Device Lett.*, vol. 38, no. 7, pp. 910–913, 2017.

[31] B. S. Kumar, M. Paul, and M. Shrivastava, “On the design challenges of drain extended FinFETs for advance SoC integration,” in *International Conference on Simulation of Semiconductor Processes and Devices*, 2017, pp. 189–192.

[32] J. Li and P. Li, “High Permittivity Dielectric LDMOS for Improved Performance,” *Energy Procedia*, vol. 12, pp. 341–347, 2011.

[33] T. Erlbacher, G. Rattmann, A. J. Bauer, and L. Frey, “Trench gate integration into planar technology for reduced on-resistance in LDMOS devices,” *Proc. Int. Symp. Power Semicond. Devices ICs*, pp. 181–184, 2010.

[34] R. Waser, *Nanoelectronics and Information Technology*, 1st ed. John Wiley & Sons, 2012.

[35] A. Hokazono *et al.*, “Steep channel & halo profiles utilizing boron-diffusion-barrier layers (Si:C) for 32 nm node and beyond,” in *Digest of Technical Papers - Symposium on VLSI Technology*, 2008, vol. 39, no. 1, pp. 112–113.

[36] Y. Cho, S. K. Kwon, H. Jung, and J. Kim, “High Performance Power MOSFETs with Strained-Si Channel,” in *17th International Symposium on Power Semiconductor Devices and ICs*, 2005, pp. 191–194.

[37] S. Sun, “Power Metal-Oxide-Semiconductor Field-Effect Transistor with Strained Silicon and Silicon Germanium Channel,” University of Central Florida, 2010.

[38] J. H. Davies, *The Physics of Low-Dimensional Semiconductors: An Introduction*, 1st ed. Cambridge University Press, 1998.

[39] R. F. Pierret, *Field effect devices*, 2nd ed. Purdue University: Addison-Wesley Publishing Company, 1990.

[40] R. W. Bower, “Field-effect device with insulated gate,” 1969.

[41] Y. Singh and R. S. Rawat, “High figure-of-merit SOI power LDMOS for power integrated circuits,” *Eng. Sci. Technol. an Int. J.*, vol. 18, no. 2, pp. 141–149, 2015.

[42] T. Mori, H. Fujii, S. Kubo, and T. Ipposhi, “Investigation into HCl improvement by a split-recessed-gate structure in a STI-based nLDMOSFET,” in *Proceedings of the International Symposium on Power Semiconductor Devices and ICs*, 2017, no. c, pp. 459–462.

[43] R. J. Mears, J. Augustin Chan Sow Fook Yiptong, M. Hytha, S. A. Krepps, and I. Dukovski, “Semiconductor Device Including A MOSFET Having A Band-Engineered Superlattice With A Semiconductor Cap Layer Providing A Channel,” 2005.

[44] Statista, “Semiconductor industry sales worldwide 1987-2019.” [Online]. Available: https://www.statista.com/statistics/266973/global-semiconductor-sales-since-1988/.

[45] B. E. Deal, “Standardized Terminology for Oxide Charges Associated with Thermally Oxidized Silicon,” *IEEE Trans. Electron Devices*, vol. 27, no. 3, pp. 606–608, 1980.

[46] H. Veendrick, *Nanometer CMOS ICs - From Basics to ASICs*, 1st ed. Springer, 2008.

[47] Synopsys, *Sentaurus^{TM} Device User Guide*, no. N-2017.09. 2017.

[48] J. Yipeng, W. Kangliang, W. Taihuan, D. Gang, and L. Xiaoyan, “Comparison of band-to-band tunneling models in Si and Si — Ge junctions,” *J. Semicond.*, vol. 34, no. 9, pp. 1–4, 2013.

[49] B. Van Zeghbroeck, “Chapter 7: MOS Field-Effect-Transistors,” in *Principles of Semiconductor Devices*, 2011.

[50] Z. L. Yue Fu, Johnny K. O. Sin, Wai Tung Ng, *Integrated Power Devices and TCAD Simulation*, 1st ed. Boca Raton: CRC Press, 2014.

[51] C. Hu, *Modern Semiconductor Devices for Integrated Circuits*, 1st ed. Prentice Hall, 2010.

[52] J. S. J. C. D. Ke, "The analysis and modeling of on-resistance in high-voltage LDMOS," in *8th International Conference on Solid-State and Integrated Circuit Technology Proceedings*, 2006, pp. 1327–1329.

[53] A. W. Ludikhuize, "Kirk effect limitations in High Voltage IC's," in *Proceedings of the 6th international Symposium on Power Semiconductor Devices & IC's*, 1994, no. 94, pp. 3–6.

[54] C. Chu *et al.*, "Investigation of voltage-dependent drift region resistance on high-voltage drain-extended MOSFETs' I-V characteristics," *Electron. Lett.*, vol. 48, no. 2, pp. 2–3, 2012.

[55] B. S. Kumar and M. Shrivastava, "Part I : On the Unification of Physics of Quasi-Saturation in LDMOS Devices," *IEEE Trans. Electron Devices*, vol. 65, no. 1, pp. 191–198, 2018.

[56] J. Mung Park, "Novel Power Devices for Smart Power Applications," Technical University Wien, 2004.

[57] B. Thidé, *Electromagnetic Field Theory*, 8th ed. Uppsala: Upsilon Books, 2004.

[58] F. Jin *et al.*, "Best-in-class LDMOS with ultra-shallow trench isolation and p-buried layer from 18V to 40V in 0.18um BCD technology," in *Proceedings of the 29th International Symposium on Power Semiconductor Devices and ICs*, 2017, pp. 295–298.

[59] H. Chou *et al.*, "0.18 um BCD Technology Platform with Best-in-Class 6V to 70V Power MOSFETs," in *ISPSD*, 2012, no. June, pp. 401–404.

[60] J. Lee *et al.*, "0.13 um modular BCD technology enable to embedding high density E2PROM, RF and Hall sensor suitable for IoT application," in *ISPSD*, 2016, pp. 419–422.

[61] K. Iwamoto *et al.*, "Advanced 300 mm 0.13 um BCD technology from 5V to 80V with highly reliable embedded Flash," in *ISPSD*, 2014, pp. 402–405.

[62] K. Shirai, K. Yonemura, K. Watanabe, and K. Kimura, "Ultra-low On-Resistance LDMOS Implementation in 0.13 um CD and BiCD Process Technologies for Analog Power IC's," in *Proceedings of the International Symposium on Power Semiconductor Devices and ICs*, 2009, pp. 77–79.

[63] H. Fujji, S. Tokumitsu, T. Mori, T. Yamashita, and T. Maruyama, "A 90nm BiCDMOS Platform Technology with 15-80V LD-MOSFETs for Automative Applications," in *ISPSD*, 2017, vol. 6, pp. 7–10.

[64] A. Chaudhry, *Fundamentals of Nanoscaled Field Effect Transistors*, 1st ed. Springer New York, 2013.

[65] B. Cheng and J. Woo, "Measurement and Modeling of the n-channel and p-channel MOSFET's Inversion Layer Mobility at Room and Low Temperature Operation," *J. Phys. IV Colloq.*, vol. 06, no. C3, pp. 1–6, 1996.

[66] C. Hamaguchi, *Basic Semiconductor Physics*, 3d edition. Osaka: Springer, 2017.

[67] S. Takagi, "Strained-Si CMOS Technology," in *Advanced Gate Stacks for High-Mobility Semiconductors*, 1st ed., Springer, 2007, pp. 1–17.

[68] L. Wang, "Quantum Mechanical Effects On MOSFET Scaling Limit," Georgia Institute of Technology, 2006.

[69] C. P. Ewels, "Density Functional Modelling of Point Defects in Semiconductors," University of Exeter, 1997.

[70] H. Takeuchi *et al.*, "Punch-Through Stop Doping Profile Control via Interstitial Trapping by Oxygen-Insertion Silicon Channel," *Electron Device Soc.*, vol. 6, pp. 481–486, 2018.

[71] J. C. Mikkelsen, "The diffusivity and solubility of oxygen in silicon," in *Materials Research Society*, 1986, vol. 59, pp. 19–30.

[72] D. Colleoni and A. Pasquarello, "Diffusion of interstitial oxygen in silicon and germanium: a hybrid functional study," *J. Phys. Condens. Matter*, vol. 28, pp. 1–7, 2016.

[73] P. A. Stolk, H. Gossmann, D. J. Eaglesham, D. C. Jacobson, and C. S. Rafferty, "Physical mechanisms of transient enhanced dopant diffusion in ion-implanted silicon," *Appl. Phys.*, vol. 81, no. 9, pp. 6031–6050, 1997.

[74] S. W. Jones, "Silicon Integrated Circuit Process Technology," in *Diffusion in Silicon*, IC Knowledge LLC, 2008, pp. 1–71.

[75] R. J. Mears *et al.*, "Punch-Through Stop Doping Profile Control via Interstitial Trapping by Oxygen-Insertion Silicon Channel," in *IEEE Electron Devices Technology and Manufacturing*, 2017, pp. 7–8.

[76] X. Zhang *et al.*, "Effects of oxygen-inserted layers on diffusion of boron, phosphorus and arsenic in silicon for ultra-shallow junction formation," *J. Appl. Phys.*, vol. 123, no. 125704, pp. 1–7, 2018.

[77] J. A. Pals, "Quantization effects in semiconductor inversion and accumulation layers," University of Eindhoven, 1972.

[78] R. J. Mears *et al.*, "Silicon Superlattice on SOI for High Mobility and Reduced Leakage," in *IEEE International SOI Conference Proceedings*, 2007, pp. 23–24.

[79] Synopsys, *Sentaurus™ Process User Guide*, no. N-2017.09. 2017.

[80] A. Ortiz-Conde, F. J. García Sánchez, J. J. Liou, A. Cerdeira, M. Estrada, and Y. Yue, "A review of recent MOSFET threshold voltage extraction methods," *Microelectron. Reliab.*, vol. 42, no. 4–5, pp. 583–596, 2002.

[81] CascadeMicrotech, "Data sheet CM300," 2016. [Online]. Available: https://www.cascademicrotech.com/files/CM300_DS.pdf. [Accessed: 10-Jul-2018].

[82] Keysight, "Data sheet Keysight B1500A semiconductor device analyzer," 2018. [Online]. Available: http://literature.cdn.keysight.com/litweb/pdf/5989-2785EN.pdf. [Accessed: 10-Jul-2018].

[83] Keysight, "Keysight EasyEXPERT Software User Guide," *Volume 1*, 2016. [Online]. Available: http://literature.cdn.keysight.com/litweb/pdf/B1540-90000.pdf.

[84] L. Stauffer, "C-V Measurement Tips, Tricks, and Traps," *Keithley Instruments, Inc.*, no. 3109, Ohio, pp. 1–14, 2011.

[85] K. Onishi *et al.*, "Improvement of Surface Carrier Mobility of HfO2 MOSFETs by High-Temperature Forming Gas Annealing," *IEEE Trans. Electron Devices*, vol. 50, no. 2, pp. 384–390, 2003.

[86] L. C. Wagner, *Failure Analysis of Integrated Circuit: Tools and Techniques*, 1st ed. Kluwer Academic Publishers, 1999.

[87] E. C. Griffith *et al.*, "Capacitance Modelling of LDMOS Transistors," in *30th European Solid-State Device Research Conference*, 2000, pp. 1–4.

[88] A. A. Brown, L. J. Neelands, and H. E. Farnsworth, "Thermionic Work Function of the (100) Face of a Tungsten Single Crystal," *J. Appl. Phys.*, vol. 21, no. 1, pp. 1–4, 1950.

[89] Synopsys, *Sentaurus™ Structure Editor User Guide*, no. N-2017.09. 2017.

[90] I. H. Ji *et al.*, "New power MOSFET employing segmented trench body contact for improving the avalanche energy," in *Proceedings of the International Symposium on Power Semiconductor Devices and ICs*, 2008, pp. 115–118.

Appendices

Appendix A: LDMOS resistance model

A.1 Flat band voltage in accumulation region

The flat band voltage (V_{FB}) is defined as the gate voltage at which there is no band bending in the semiconductor and as a result zero net space charge in this region. In a LDMOS device, the gate covers the channel and the accumulation region (Figure 5 (left)), which both have a separate V_{FB}. Since the V_{FB} of the accumulation layer basically acts as the V_T of the channel (formula wise), it is of particular interest for proper modeling of the R^*_{acc} (Equation (9)). One difficulty however, is that unlike the V_T, the V_{FB} (from the channel and/or accumulation region) cannot be determined from simple electrical measurements. Theoretically, V_{FB} can be determined from the ϕ_{ms}, C_{ox} and Q'_{ss} as is shown in Equation (2). Typically, the oxide charge Q'_{ss} only causes a small shift in V_{FB}, such that for simplicity it can be assumed that $V_{FB} \approx \phi_{ms}$ [11]. Considering the case of a n-type LDMOS, the work function difference between the gate and the n-type doped semiconductor in the accumulation region can be determined according to the following equation: [11]

$$\phi_{ms} = \phi_m - (\chi + \frac{E_g}{2e} - \frac{k_B T}{e} \ln\left(\frac{N_d}{n_i}\right)) \tag{30}$$

Here ϕ_m is the gate work function [eV], χ the electron affinity in silicon [eV], E_g the bandgap in silicon (1.12 eV), N_d the donor concentration [cm^{-3}] and n_i the intrinsic carrier concentration in silicon (1.5x10^{15} cm^{-3}). Moreover, for a n-type LDMOS, the gate is typically made of n$^+$-polysilicon for which applies that $\phi_m \approx \chi \approx 4.05$ eV [51]. For donor doping concentrations in the range from 10^{16} - 10^{18} cm^{-3}, ϕ_{ms} then will lie in the range from -0.21V till -0.09V. In order to verify this, simulations were performed with various (n-type) doping concentrations in the accumulation region (with a n$^+$-polysilicon gate) with zero bias at all the terminals. Then, by plotting the conduction band energy through the gate, oxide and silicon, V_{FB} (=ϕ_{ms}) can be determined by summing the drop over the oxide and over the silicon [11]. An example of such a simulation (where N_D ≈3x10^{17} cm^{-3}) is shown in the figure below:

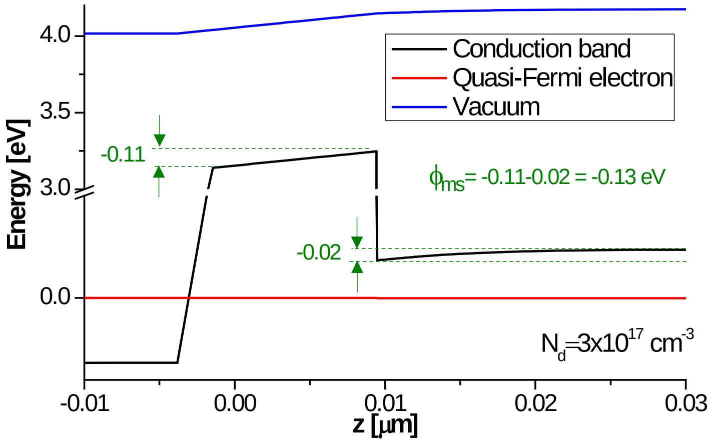

Figure 65: Conduction band, quasi-Fermi electron and vacuum energy through the (n$^+$-polysilicon) gate, oxide and (n-type doped ~3x10^{17} cm^{-3}) silicon with all terminals at 0V. The V_{FB} (=ϕ_{ms}) can be calculated by summing the voltage drop over the oxide and over the band bending in the silicon [11].

In the figure below, the results of the simulations are shown as well as the calculated ϕ_{ms} (Equation (30)) and ϕ_{ms} data extracted from Baliga [14].

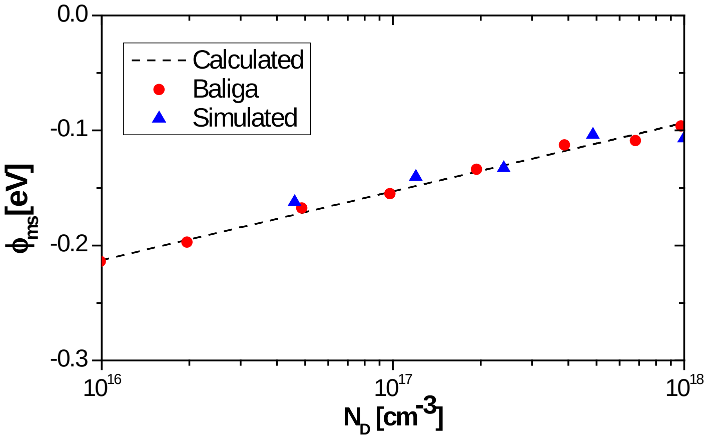

Figure 66: The metal-semiconductor work function difference for a n^+-polysilicon gate on n-type doped silicon from calculation, simulation and Baliga [14]. *The calculated ϕ_{ms} is in good agreement with the simulations and the data from Baliga* [14].

From this figure, it can be seen that the calculated ϕ_{ms} is in good agreement with the simulations and the data from Baliga [14].

A.2 Resistance separation method

In the expression for the R_{ON} set up for a LDMOS device (Equation (11)), the first two terms depend on V_{GS} while the other terms are constant. Therefore, by introducing some fitting parameters, this expression can be rewritten as:

$$R_{ON} = \frac{1}{\alpha(V_{GS} - V_T)} + \frac{1}{\beta(V_{GS} - V_{FB})} + R_{res} \tag{31}$$

By fitting this equation to a $\mathrm{R_{ON}}$-V_{GS} curve for $V_{GS} > V_T$ (where $\mathrm{R_{ON}}$=$0.1/I_{D,lin}$), the parameters α, β and R_{res} can be determined. Subsequently, α can be used to calculate R_{ch}, β can be used to calculate R_{acc} and R_{res} equals R_{drift}+R_{sd}.

Since this approach requires the fitting of three parameters, one could imagine that there are many different fitting solutions, who may not all be physical accurate. This would make the fitting procedure very sensitive to the accuracy of the input parameters, the range of fitting and the starting condition. In order to reduce the number of possible solutions, it would be beneficial if multiple $\mathrm{R_{ON}}$-V_{GS} curves could be obtained in which one or multiple input parameters are varied. A novel way to do this, is changing the V_T by biasing the body/substrate. The shift in threshold voltage as a function of the substrate bias or source-to-body bias ($V_{SB} = V_S - V_B$) is given by: [11]

$$\Delta V_T = \frac{\sqrt{2e\epsilon_s N_a}}{C_{ox}}\left(\sqrt{2\phi_{fp} + V_{SB}} - \sqrt{2\phi_{fp}}\right) \tag{32}$$

Mathematically, this expression only gives real solutions when the two following criterions are met:

$$2\phi_{fp} + V_{SB} > 0 \quad \& \quad \phi_{fp} > 0 \tag{33}$$

Filling in the expression for ϕ_{fp} (Equation (2)), these criteria result in the following equation:

$$V_{SB} > \frac{2k_BT}{e}\ln\left(\frac{n_i}{N_a}\right) \quad where\ N_a > n_i \tag{34}$$

For a typical acceptor doping concentration of 10^{17} cm^{-3} at room temperature, this results in that V_{SB} should be bigger than ~-0.8V. Physically, this result could have been anticipated since biasing the body-to-source junction above the junction voltage (which is about 0.6V for silicon [46]) puts the junction in forward bias.

The following approach is then used to separate the resistances in a LDMOS device:

1. Measure the I_D-V_{GS} curves for several V_{SB} >-0.6V (e.g. -0.5, 0 and 0.5V).
2. Extract the $V_{T,gm}$ from each of these curves.
3. Determine or estimate V_{FB} of the accumulation region. For a n-type LDMOS device with a n$^+$-polysilicon gate Figure 66 can be used.
4. Calculate the R$_{\mathrm{ON}}$-V_{GS} curves by R$_{\mathrm{ON}}$=0.1/I_D.
5. Set up separate fits for the R$_{\mathrm{ON}}$-V_{GS} curves according to the fitting equation (Equation (31)). Make sure the starting condition of the fitting parameters is a good initial guess. The range of fitting should be such that the steep diverging part for $V_{GS} \rightarrow V_T$ is excluded, since this is hard to fit. A good guideline is fitting each curve for $V_{GS} > V_T + 0.5$.
6. Calculate the residual sum of squares for all the fits and minimize this value iteratively by a solver in which the fitting parameters are used as input. The number of iterations should be such that the solution is converged.
7. Finally, check whether the solution make sense, otherwise re-evaluate the starting conditions and fitting domain.

Appendix B: Derivations

B.1 1D breakdown approximation analysis

For a one-sided p$^+$n or (or pn$^-$) junction with applied reverse bias voltage V_R, the depletion width can be determined by taking the limit of $N_d \to 0$ for Equation (1b), which results in:

$$x_n = W_D = \left(\frac{2\epsilon_s V_R}{eN_d}\right)^{\frac{1}{2}} \quad \& \quad x_p = 0 \tag{35}$$

Here it is assumed that $V_R \gg V_{bi}$, such that $V_R + V_{bi} \approx V_R$. The electric field extending in the n-region can now be written as (Equation (1c)):

$$E(0 \leq x \leq W_D) = \frac{-eN_d}{\epsilon_s}(W_d - x) = \frac{-eN_d}{\epsilon_s}\left(\left(\frac{2\epsilon_s V_R}{eN_d}\right)^{\frac{1}{2}} - x\right) \tag{36}$$

The criterium for avalanche breakdown is given by Equation (5) (with M=1) and together with Fulop's approximation [14] it can be expressed as:

$$\int_0^{W_D} 1.8 \cdot 10^{-35} E^7 dx = 1 \tag{37}$$

Filling in Equations (35) & (36) and solving for V_R then gives the breakdown voltage as a function of the doping concentration in the n-type region: [14]

$$V_R = BV = 5.34 \cdot 10^{13} \cdot N_d^{-\frac{3}{4}} \tag{38}$$

The corresponding depletion width can be calculated by substituting Equation (38) in Equation (35): [14]

$$W_{D,bd} = 2.67 \cdot 10^{10} \cdot N_d^{-\frac{7}{8}} \tag{39}$$

B.2 Ideal 1D trade-off drift region

In this section, the ideal trade-off between the specific drift resistance and the off-state breakdown voltage (due to avalanche breakdown) for a 1D LDMOS device with uniform doping is derived. Since the analysis is purely 1D, the influence of the gate is neglected such that L_{acc} can be set to zero. Again, here the assumption is made of a one-sided p$^+$n-junction (or pn$^-$) such that when a reverse bias voltage is applied, the drift region will almost entirely accommodate the voltage drop. The trade-off can then be derived for the NPT- and PT case, which are sketched in the figure below:

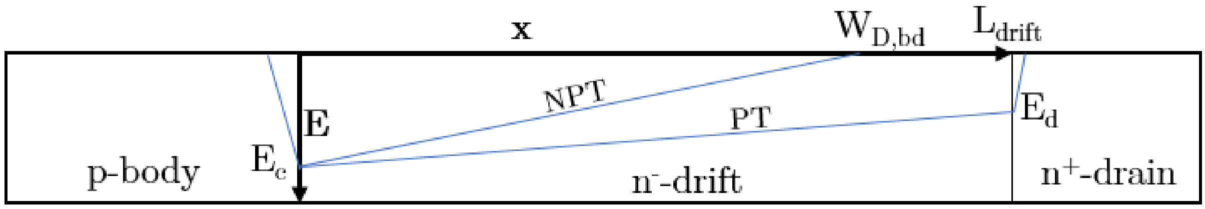

Figure 67: Schematic sketch for deriving the ideal trade-off between the specific drift resistance and the off-state breakdown voltage for a 1D LDMOS device. Two different situations can be considered here: the non-punch through condition (NPT) and the punch through condition (PT).

For each situation, the BV can be expressed as follows:

$$\mathit{NPT}\colon\, BV = \frac{E_c}{2} W_{D,bd} \quad \vee \quad \mathit{PT}\colon\, BV = \frac{E_c + E_d}{2} L_{drift} \tag{40}$$

Moreover, from Equation (1c) it is known that the slope of the electric field is proportional to eN_d/ϵ_s, such that the following expressions can be derived:

$$\mathit{NPT}\colon\, E_c = \frac{eN_d}{\epsilon_s} W_{D,bd} \quad \vee \quad \mathit{PT}\colon\, E_d = E_c - \frac{eN_d}{\epsilon_s} L_{drift} \tag{41}$$

The specific drift resistance can be obtained by multiplying Equation (10) by WL_{drift}, which results in:

$$R_{drift}A_{drift} = R_{drift} \cdot WL_{drift} = \frac{1}{eN_d \mu_n} \left(\frac{L^2_{drift}}{d_{eff}} \right) \tag{42}$$

Combining Equations (40) and (41) to solve for eN_d and filling this in for Equation (42) then gives:

$$\mathit{NPT}\colon\, R_{drift}A_{drift} = \frac{L^2_{drift} W^2_{D,bd}}{2BV \epsilon_s \mu_n d_{eff}} \quad \vee \quad \mathit{PT}\colon\, R_{drift}A_{drift} = \frac{L^4_{drift}}{2(E_c L_{drift} - BV)\epsilon_s \mu_n d_{eff}} \tag{43}$$

For the NPT condition, the trade-off is ideal when $L_{drift} = W_{D,bd} = 2BV/E_c$ since this maximizes the BV at a fixed R_{drift}. For the PT condition, the optimal drift region length can be found by solving $\frac{\partial R_{drift} A}{\partial L_{drift}} = 0$, which results in $L_{drift} = 4BV/3E_c$ [1]. Filling in these considerations subsequently gives:

$$\mathit{NPT}\colon\, R_{drift}A_{drift} = \frac{8BV^3}{\epsilon_s \mu_n d_{eff} E_c^4} \quad \vee \quad \mathit{PT}\colon\, R_{drift}A_{drift} = \frac{128BV^3}{27\epsilon_s \mu_n d_{eff} E_c^4} \tag{44}$$

In the case of silicon power MOSFET's, the doping concentration in the drift region is sufficiently low, such that the mobility typically is assumed to be constant [14]. The E_c and BV however, are related to the doping concentration in the drift region according to Equation (15) and Equation (38), which are typically used for both conditions [1], [14]. Combining these equations for the critical field results in:

$$E_c = 7.78 \cdot 10^5 BV^{-\frac{1}{6}} \tag{45}$$

Substituting this expression in the Equations in (44) then yields the ideal trade-off for the drift region of 1D silicon LDMOS devices:

$$\begin{gathered} R_{drift}A_{drift}\ [\Omega \cdot cm^2] = C_{1D} \cdot BV^{\frac{11}{3}} \\ \text{Where} \quad C_{1D}(\mathit{NPT}) = \tfrac{1.55 \cdot 10^{-14}}{d_{eff}} \quad \vee \quad C_{1D}(\mathit{PT}) = \tfrac{9.18 \cdot 10^{-15}}{d_{eff}} \end{gathered} \tag{46}$$

In calculating these constants, 1360 cm^{2}V^{-1} s^{-1} is used for the electron mobility which is valid in the low doping regime [14]. From these constants, it can be seen that the PT condition provides the best trade-off. Therefore, the $C_{1D}(PT)$ constant is used to describe the ideal trade-off for the drift region of 1D silicon LDMOS devices.

B.3 Ideal RESURF trade-off drift region

In this section, the ideal trade-off between the specific drift resistance and the off-state breakdown voltage (due to avalanche breakdown) for a 2D device with uniform doping under ideal RESURF conditions is derived. The same situation as the previous section (B.2) is assumed, only now the electric field in the current direction (x) is uniform and fully extended to the drain. This situation is sketched in the figure below:

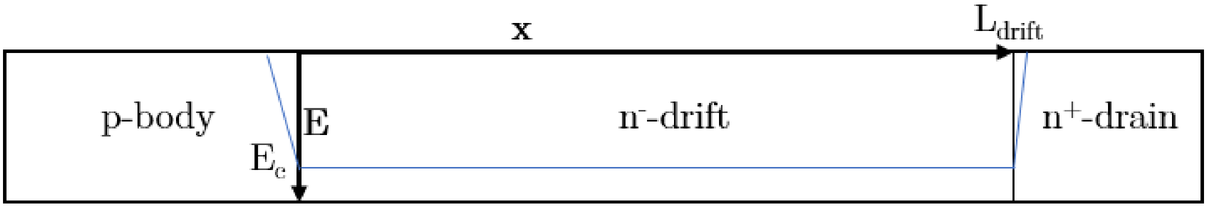

Figure 68: Schematic sketch for deriving the ideal trade-off between the specific drift resistance and the off-state breakdown voltage for a 2D device under ideal RESURF conditions.

Using Equation (5) (with M=1) together with Fulop's approximation [14] and setting $W_D \rightarrow L_{drift}$, then gives for the onset of avalanche breakdown:

$$\int_0^{L_{drift}} 1.8 \cdot 10^{-35} E^7 dx = 1 \;\rightarrow\; 1.8 \cdot 10^{-35} E_c^7 L_{drift} = 1$$

Using $BV = E_c L_{drift}$, this can be rewritten as:

$$L_{drift} = (1.8 \cdot 10^{-35})^{\frac{1}{6}} BV^{\frac{7}{6}} \tag{47}$$

This equation can subsequently be substituted in the specific drift resistance (Equation (42)), yielding a relation between the specific drift resistance and the off-state breakdown voltage:

$$R_{drift} A_{drift}\;[\Omega \cdot cm^2] = \frac{(1.8 \cdot 10^{-35})^{\frac{1}{3}}}{e N_d \mu_n d_{eff}} BV^{\frac{7}{3}} = C_{2D} \cdot BV^{\frac{7}{3}} \tag{48}$$

Note that in the pre-factor, in contrast to the 1D trade-off, now also the doping concentration is contained. This is because the ideal doping concentration is dependent on the RESURF application and can thus not be generalized for ideal RESURF devices. Moreover, since with RESURF applications typically higher doping concentrations are used in the drift region, it is not always legit to use 1360 cm^{2}V^{-1} s^{-1} for the electron mobility valid in the low doping regime [14]. Using an empirical relation for μ_n as a function of N_d [14], the pre-factor can then be expressed as:

$$C_{2D} = \frac{(1.8 \cdot 10^{-35})^{\frac{1}{3}}}{e N_d \mu_n(N_d) d_{eff}} = \frac{1.64 \cdot 10^7}{d_{eff} N_d}\left(\frac{1}{\mu_n(N_d)}\right) = \frac{1.64 \cdot 10^7}{d_{eff} N_d}\left(\frac{3.75 \cdot 10^{15} + N_d^{0.91}}{5.10 \cdot 10^{18} + 92 N_d^{0.91}}\right) \tag{49}$$

Appendix C: Supporting simulations CFP

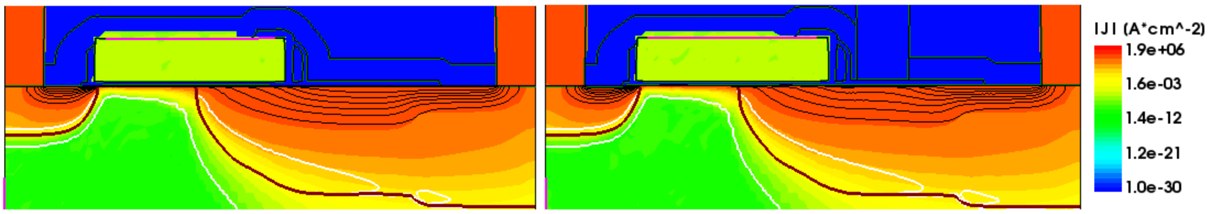

Figure 69: $I_D - V_{GS}$ simulations of the default FA- (left) and CFP device (right) at V_{GS}=5V and V_{DS}=0.1V. The color indicates the absolute current density (|J|) and the black lines are current potential lines. The default CFP device exhibits slightly larger $I_{D,lin}$ because of the additional accumulation below the CFP.

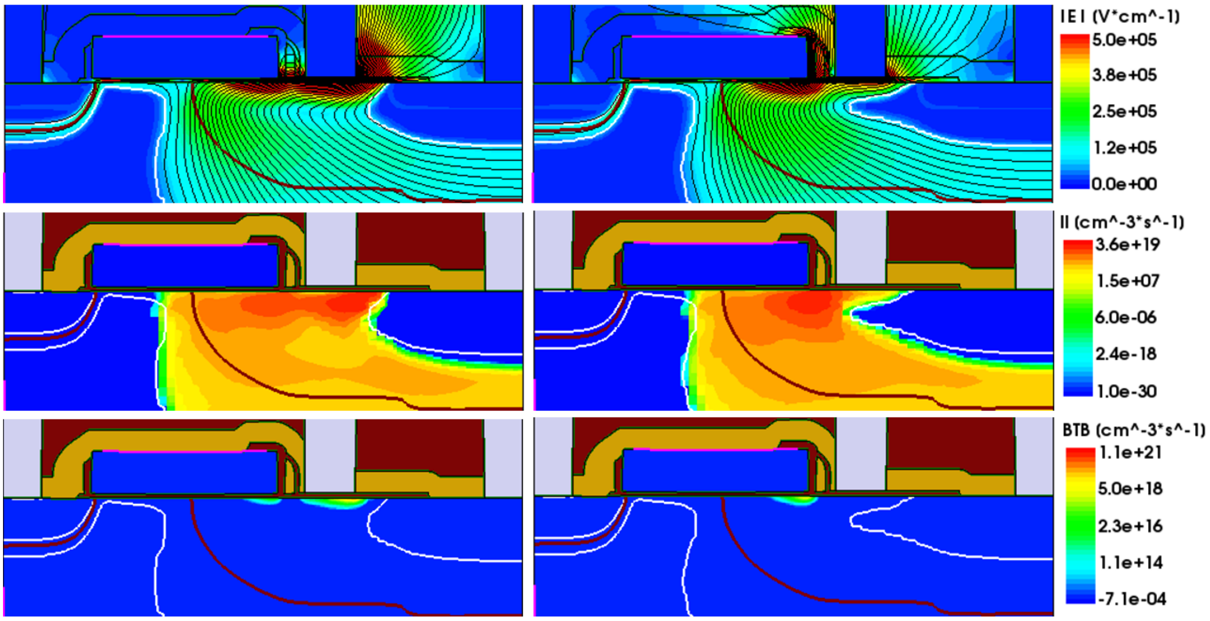

Figure 70: Off-state breakdown simulations (at $V_{DS} \approx 12V$) of the default CFP device with a FP bias of 0V (left) and 7V (right). Top: Magnitude of electric field (|E|) and equipotential lines (black lines). Middle: Impact ionization (II). Bottom: Band-to-band generation (BTB). Biasing the CFP at 7V reduces the field peak below the CFP at the cost of the field peak below the gate edge and reduces the depletion width (indicated by white line) in the drift region. This reduces the impact ionization as well as the band-to-band generation.

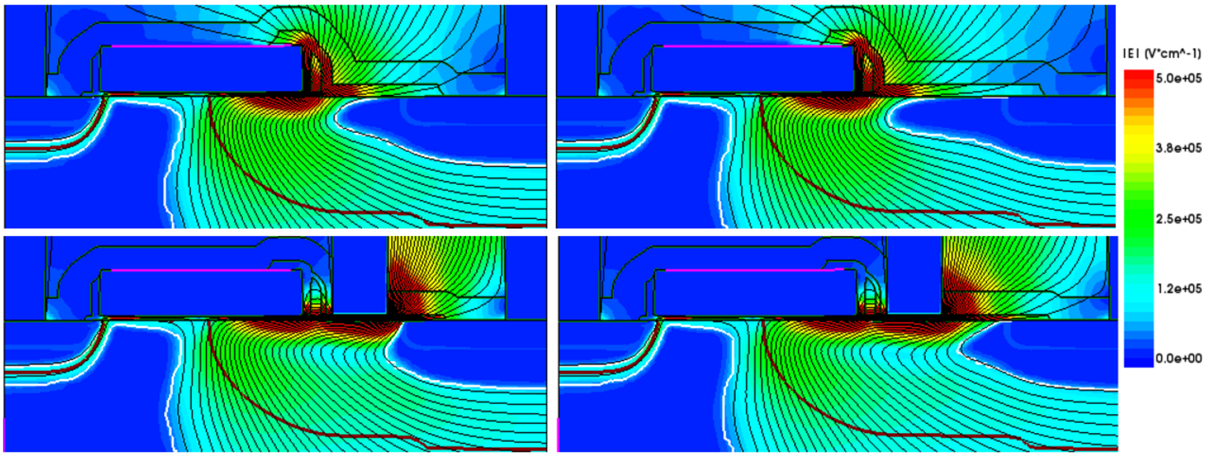

Figure 71: Off-state breakdown simulations (at $V_{DS} \approx 12V$) of the FA- (top) and CFP device (bottom) with $L_{drift} = 0.32$ µm (left) and $L_{drift} = 0.50$ µm (right). Here the magnitude of the electric field is shown as well as the equipotential lines. For the FA device, it can be seen that increasing L_{drift} basically has no use since the depleted region (indicated by white line) does not scale accordingly. For the CFP device however, the depletion width is increased with increasing L_{drift}. Therefore, the same voltage is supported over a longer region such that better RESURF conditions apply resulting in a higher BV.

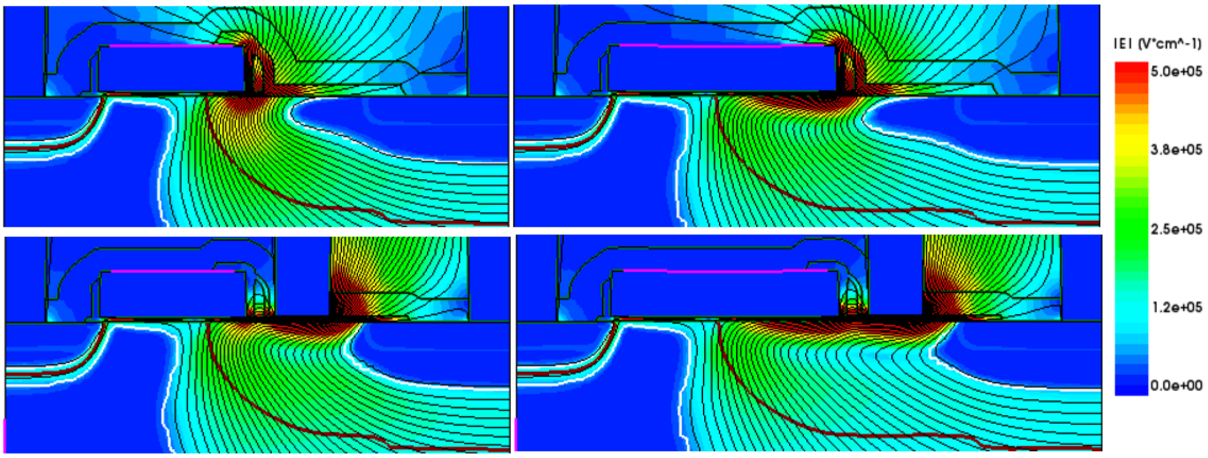

Figure 72: Off-state breakdown simulations (at $V_{DS} \approx 12V$) of the FA- (top) and CFP device (bottom) with L_{acc}=0.1 µm (left) and L_{acc}=0.4 µm (right). Here the magnitude of the electric field is shown as well as the equipotential lines. For both devices, it can be seen that increasing L_{acc} results in a broader field peak at the gate edge. For the FA device, this results in a significant increase in the band-to-band generation and thereby in the leakage current. For the CFP device, the band-to-band tunneling is dominant below the CFP, such that the broader peak at the gate edge has less influence on the leakage current.

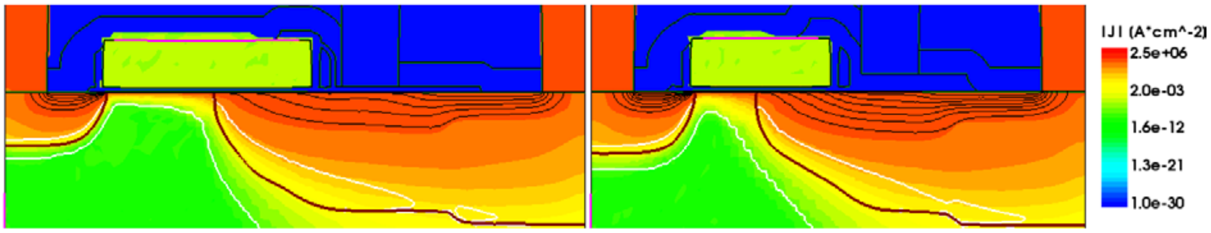

Figure 73: $I_D - V_{GS}$ simulations of the default CFP device (left) and the version optimized for 12V application (right) at V_{GS}=5V and V_{DS}=0.1V. The color indicates the absolute current density (|J|) and the black lines are current potential lines. The optimized device exhibits larger $I_{D,lin}$ because of the lower doped channel and the decreased dimensions.

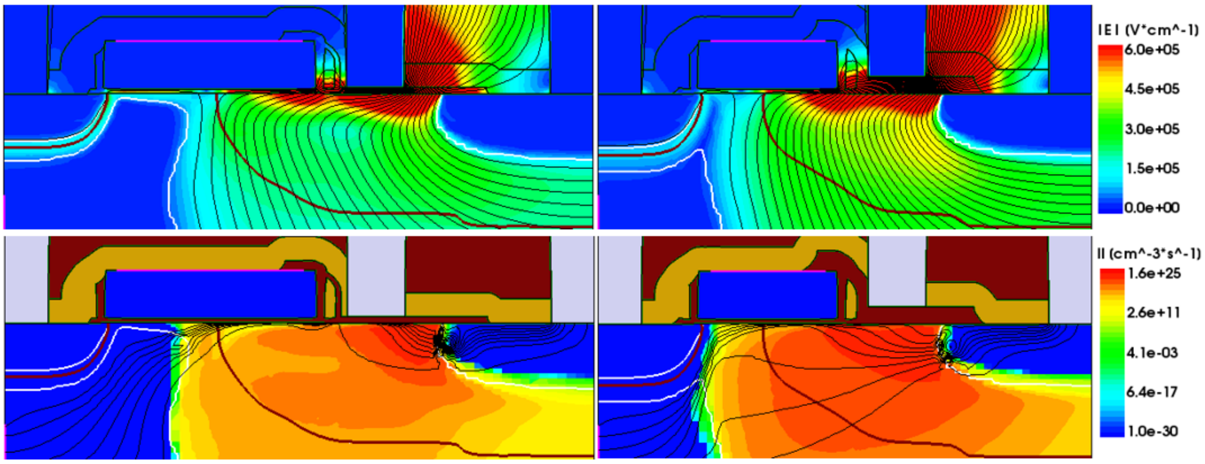

Figure 74: Off-state breakdown simulations of the default CFP device (left) and the version optimized for 12V application (right) at I_D=1x10^{-8} A/µm. Top: Magnitude of electric field (|E|) and equipotential lines (black lines). Bottom: Impact ionization (II) and current potential lines (black lines). In the optimized version, the electric field is distributed more uniformly in the drain-extension. As a consequence, a larger potential can be built-up (~5V) before avalanching breakdown sets in.

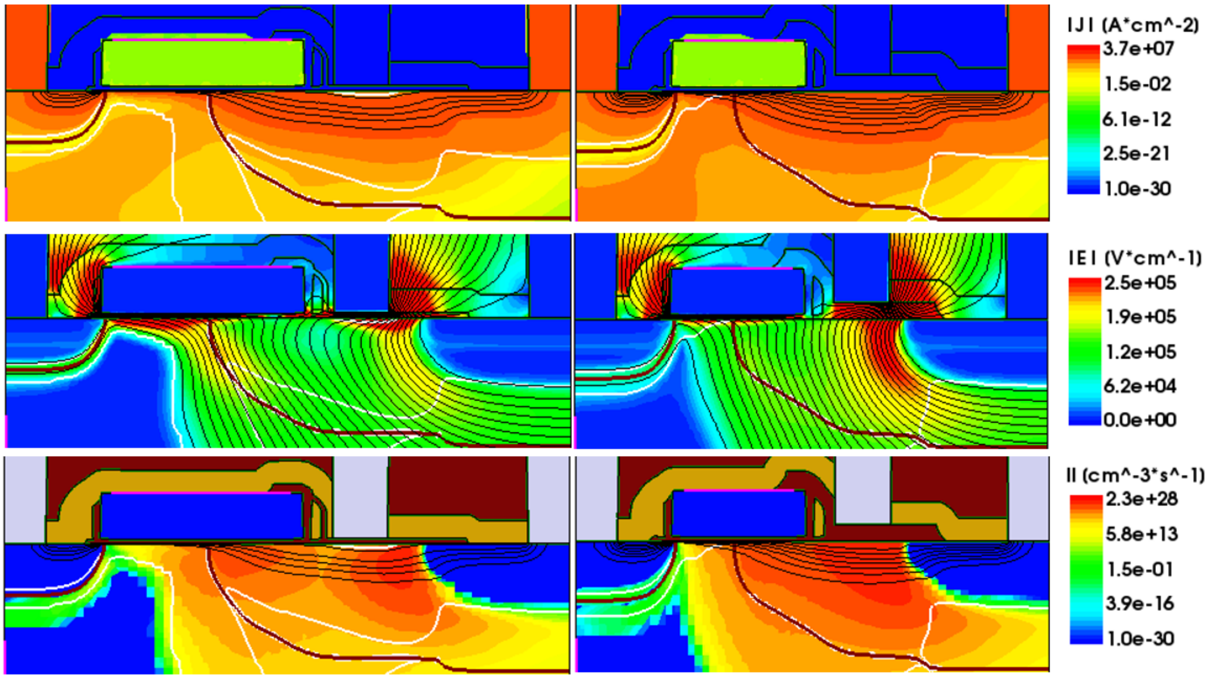

Figure 75: On-state simulations at V_{DS} ≈12V and V_{GS} =5V for the default CFP device (left) and the CFP device optimized for 12V application (right). Top: Magnitude of current density (|J|) and current potential lines (black lines). Middle: Magnitude of electric field (|E|) and equipotential lines (black lines). Bottom: Impact ionization (II) and current potential lines (black lines). For the 12V optimized CFP device, it is evident that the two centers for II merge into one large area of increased II, resulting in lower BV_{ON}. The reason for this is the higher current density (lower R_{ON}) and higher electric field below the CFP mainly due to the reduced lateral dimensions.