Jürgen Rainer Paust, BSc

Early life failure mechanisms detection in power MOSFETs by means of thermal burn-in

MASTER'S THESIS

to achieve the university degree of

Diplom-Ingenieur

Master's degree programme: Technical Physics

submitted to

Graz University of Technology

Supervisor

Univ.-Prof. Peter Hadley, Ph.D.

Institute of Solid State Physics

Graz, July 2023

Acknowledgements

At this point I want to thank all the people who have supported me during the writing of this thesis.

First and foremost my professor at university, Peter Hadley, and my mentor at Infineon Technologies, Thomas Kuczmik. Without their valuable advice and mentoring I would not have been able to manage this thesis as well as I have.

Furthermore I want to thank Infineon Technologies and all the people that work there, especially Marian Peschetz, Christian Kovacs and all my product engineering colleagues, for all the time, patience and resources they spent in order to help me.

The last thanks goes out to my family and friends. It was their love and support that gave me the opportunity to come this far in the first place.

Abstract

As a part of the development process of new transistors, the devices need a thorough investigation of their reliability. One part of this is the electrical testing and assessing if they operate in the manners wished for. If this initial testing proved to be successful, the next step is to determine if the devices will continue to work properly throughout their specified lifetime. Since this time span can be several years, the devices can undergo an aging process by electrical and thermal stressing. The process which does this is called burn-in. Hereby, the devices will be operated with overvoltage as well as in over temperature conditions, in order to see, if these conditions will cause the device to fail. During the course of this thesis, 50 chips, with several MOSFETs on them have experienced these sorts of tests and the resulting data was evaluated. Interpreting the data was achieved by having a closer look on the physics behind the failure mechanisms the devices are prone to and finding the relevant parameters which indicate, whether or not, a fail was activated during burn-in. It has been seen, that none of the devices on the 50 chips have failed and that the manufacturing process is therefore optimized enough, to guarantee reliability of the devices (as far as the gathered data is concerned). Another thing that has been seen is, that the methods used to evaluate the data would have successfully screened out any malfunctioning devices. During the experiments conducted, it was shown, that if the gate oxide breaks down, the resulting gate leakage current increases by orders of magnitudes, which would have been screened by the limits applied to the test data. Therefore it can be assumed, that no such failure had occurred in the new power MOSFETs of Infineon Technologies and that the methods used to evaluate the data are working.

Kurzfassung

Im Rahmen des Entwicklungsprozesses neuer Transistoren müssen die Halbleiterbauteile gründlich auf ihre Zuverlässigkeit untersucht werden. Ein Teil davon ist die elektrische Prüfung und die Beurteilung, ob sie in der gewünschten Weise funktionieren. Wenn diese ersten Tests erfolgreich verlaufen sind, muss im nächsten Schritt festgestellt werden, ob die Halbleiterbauteile während ihrer gesamten vorgesehenen Lebensdauer ordnungsgemäß funktionieren werden. Da diese Zeitspanne mehrere Jahre betragen kann, können die Halbleiterbauteile durch elektrische und thermische Beanspruchung einem beschleunigten Alterungsprozess unterzogen werden. Der Prozess, der dies bewirkt, wird Burn-in genannt. Dabei werden die Halbleiterbauteile sowohl mit Überspannung als auch unter Übertemperaturbedingungen betrieben, um zu sehen, ob diese Bedingungen zu einem Ausfall der Halbleiterbauteile führen. Im Rahmen dieser Arbeit wurden 50 Chips, auf denen sich mehrere MOSFETs befanden, derartigen Tests unterzogen und die daraus resultierenden Daten ausgewertet. Die Interpretation der Daten erfolgte durch eine genauere Betrachtung der Physik hinter den Fehlermechanismen, für die die Bauelemente anfällig sind und durch die Ermittlung der relevanten Parameter, die anzeigen, ob ein Ausfall während des Burn-in aktiviert wurde oder nicht. Es hat sich gezeigt, dass keines der Bauteile auf den 50 Chips ausgefallen ist und dass der Herstellungsprozess daher ausreichend optimiert ist, um die Zuverlässigkeit der Bauteile zu gewährleisten (soweit die gesammelten Daten dies zeigen können). Außerdem hat sich gezeigt, dass die zur Auswertung der Daten verwendeten Methoden erfolgreich alle fehlerhaften Geräte aussortiert hätten. Bei den durchgeführten Experimenten hat sich gezeigt, dass bei einem Ausfall des Gate-Oxids der resultierende Gate-Leckstrom um Größenordnungen ansteigt, was von den auf die Testdaten angewandten Limits gescreened worden wäre. Es kann daher davon ausgegangen werden, dass bei den neuen Leistungs-MOSFETs von Infineon Technologies kein solcher Ausfall aufgetreten ist und dass die zur Auswertung der Daten verwendeten Methoden funktionieren.

Contents

1 Introduction

In the semiconductor industry, developing new devices like transistors is a continuous journey to improve the technological know-how and device portfolio. However, when a new device is going to be used on a microchip, initial characterization of the device in order to guarantee its functionality, is one core aspect of the development process. This characterization is done via electrical testing. The newest power metal oxide semiconductor field effect transistors (power MOSFETs) of Infineon Technologies have yet to undergo this development phase.

Despite the fact that the device initially is functioning as desired, one major challenge for manufacturers is reliable operation over a long period of time. One method to determine the reliability is a thermal burn-in, which simulates an aging process [1]. By filtering out the devices that change their behaviour with age in an unwanted way, the company can assure reliability of the devices. Such unwanted aging effects can occur due to bad design or due to defects in the device, which were induced during production.

The aim is therefore to optimize the fabrication process of the newly developed power MOSFETs to a point where the burn-in is not necessary anymore. In order to complete this task one needs to fully understand the devices, their limitations and the mechanisms that cause the devices to fail. This thesis aims to answer a number of questions:

- How do the devices work?
- What are the most prominent failure mechanisms present?
- How do these mechanisms affect the devices?
- How can bad devices be efficiently identified?
- How good is the theory of failure mechanisms matching the physically observed failures?

The way to do this is to conduct a burn-in study in which a certain amount of devices are initially tested, then they undergo the aging process in a burn-in oven and then they are tested again to see if and how the device's characterization parameters change. The scope of the thesis lies more at the beginning of the study, including setting up the screening methods and understanding the behaviour of the devices for the rest of the study instead of the evaluation of the data of the whole study.

From here on the thesis will start with chapter two which is about a burn-in reference device (BIRD). This is the chip the newly developed MOSFETs are on. In this chapter an overview of the schematics and the layout, as well as a short introduction about the actual devices that are being tested is given. To properly understand these new devices, the theory about MOSFETs is recapitulated and the most important equations to understand their behaviour are discussed in chapter three. It starts with the MOS capacitor and leads to the description of a regular lateral MOSFET and what a drain extension is. Half of the transistors on the BIRD are double diffused power MOSFETs (DMOS), so there is also a sub chapter dedicated to them.

Since the main emphasis of the thesis is on the failures that can arise during the burn-in process and how they are affecting the transistors, a discussion about the theory behind the failure mechanisms is included in chapter four. It starts with a brief overview about what burn-in is and why it is mandatory to fulfill the industry standards. Furthermore it discusses how defects and failure mechanisms affect the devices during burn-in and how some of the most probable failure mechanisms in MOSFETs are described theoretically.

In the chapter about the electrical characterization of the BIRD, the test data before the burn-in will be presented. Is is evaluated in agreement with the theory of the MOSFETs, in order to derive screening limits for the production. Only if the devices are within these limits, they pass on to the next step, which is the burn-in process. The following chapter six presents the burn-in process itself and the data, to be able to identify any activated defects or design flaws taken from the devices that have already been burnt and were tested again. The data is evaluated by using screening methods which rely on the device's limitations. The chapter ends with a sub chapter about what happens if those limits are exceeded. To demonstrate this on the actual transistors, a few of the chips are exposed to such extreme conditions in the lab, in order to generate certain failures.

The purposefully generated failure or any eventual failure that was activated during BI, are examined in the lab, to see what caused the failure and if it can be traced back to a root cause. The method will include a description of the failure, a failure analysis and a qualitative assessment of the failure mechanisms with the theory introduced in chapter four. In the conclusion, the whole method is shortly summarized, an evaluation of the devices is done and the limitations of this thesis, as well as an outlook for further investigation is given.

2 Burn-in reference device (BIRD)

In order to characterize the newly developed semiconductor devices, it is necessary to test them electrically. The devices, which are mostly power MOSFETs are integrated in simple circuits on a chip. This chip is called a burn-in reference device (BIRD) and serves the purpose of making the testing as simple and direct as possible. A big advantage of the BIRD is that whenever a new semiconductor device is developed, the underlying layout and architecture of the chip can be reused with only small adaptations.

In this chapter the BIRD is being examined. At first the schematics are shown and one can see the semiconductor devices of interest, together with their respective contact pins. From those the advantages of BIRD over other chips that undergo burn-in can be formulated. Furthermore, the digital part is mentioned, followed by a brief discussion about what its functionality and structure are.

In the second part of this chapter the semiconductor devices are further explained. They are three MOSFETs, three power MOSFETs and one MOS capacitor. How to distinguish these devices and what their unique properties are is also included. After this chapter the structure of the BIRD and why it is beneficial for testing the devices, as well as what these devices exactly are, should be established.

2.1 BIRD schematics

The BIRD has in total 48 contact pins, most of which are connected to the MOSFETs and the MOS capacitors. To be precise, among the seven devices there are five n-channel MOSFETs (NMOS), one p-channel MOSFET (PMOS) and one MOS capacitor (MOScap). Every device is on the chip twice, with the difference between the two being the different device areas which they are occupying. The reason for this is that upon being tested and characterized, half of them are being stressed with a voltage greater than their specified operating voltage and the other half is not. This allows to investigate if the additional increase in voltage can activate failures in the respective devices [2]. However, in some chips it might be the case that the devices can not be stressed directly during testing. So in order to compare the performance of previously stressed and unstressed ones during BI, the amount of devices was doubled. The actual dimensions are determined by industry and company standards, which are of no particular relevance for the investigations done in this thesis.

Each NMOS has three contacts (enable, gate, drain), the PMOS have four contacts (enable, gate, drain, source) and the MOScaps each have one gate contact. This makes a total of 40 contacts for the devices. Seven of the remaining eight pins are used for the digital part of the chip. The final contact is connected to ground. The chip is mounted in a VQFN-48 package to accommodate the contacting needs and to have a cost efficient and reliable solution available.

The NMOS transistors are switched on, when a gate voltage is supplied, as well as when enable is on high, that means when the enable line is also supplied with a voltage. Behind the enable line is a Schmitt-trigger and up to three inverter structures with successive larger transistors attached. The reason behind this is that the Schmitt-trigger gives a steep voltage ramp up curve and the inverters increase the power that reaches the large NMOS device under test. Therefore when a gate voltage is applied without enable being on high, the inverters prevent the transistor from being switched on and when enable switches to high, a fast and steep voltage signal reaches the transistor. For the PMOS enable needs to be on low.

In figure 1, a schematic overview of the structure of the chip is given. It shows the contact pins which are numbered from 1 to 48. The three pins each NMOS requires are grouped together and made visible with green and red coloring. The same for the PMOS and the MOScaps, the pins for the digital part, however, are marked in blue. From each pin group contact wires lead to the devices, which are drawn as red and yellow squares, as well as to the digital part. The two colors the devices are drawn with indicate the two different device sizes. Yellow devices are the stressed ones and red are the unstressed ones. The ground wires are also drawn, leading from each device to the black square in the center, which is directly connected to pin number 17, the ground pin.

This figure is only for demonstration purposes, the actual circuitry of the whole chip is more complicated. It includes electro static discharge (ESD) structures on every device (it protects them from high voltages coming from their surroundings) and a digital part, which is also made up of several circuits and transistors itself. The digital part serves the purpose of storing basic data like a unique chip-ID. However, these structures and circuits are not described in further detail, since their design is of no relevance to the electrical characterization of the new devices. The reason for that is that regarding the measurement data, these structures are either completely disregarded, like in the case of the digital part or their effect can be included via a simple factor that shifts the data, like in the case of the ESD structures. If the latter is the case, it will be mentioned in the respective section.

At this point it becomes apparent, that due to its design and easy accessibility of the devices, the BIRD is a good approach to conduct burn-in studies with. Whenever a failure is detected, the failing part can be easily traced back and identified because of the chip-ID and because the devices all have their own contact pins, which would not be the case if they were integrated in larger circuits.

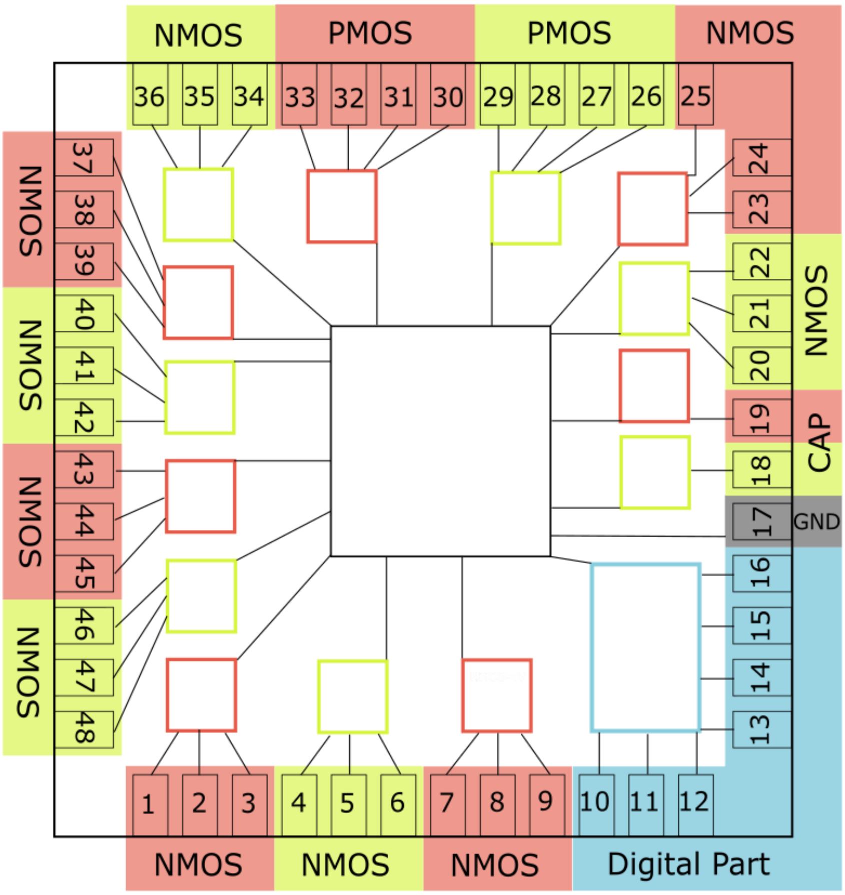


Figure 1: Toplevel schematics of the BIRD with numbering of the contacts and information about which device is contacted by which pin. The blue highlighted area is the digital part, the red and green framed parts are the seven devices that are to be investigated.

2.2 BIRD semiconductor devices

Now that the general layout of the BIRD has been introduced, the focus can be put on the actual semiconductor devices and some of their physical properties and differences. Three of the seven devices on the BIRD are conventional lateral MOSFETs, three of them are double-diffused MOSFETs (DMOS) and one of them is a MOS capacitance. However, it has to be remembered, that all of these devices are on the chip twice.

To further distinguish the devices, the most relevant differences are summarized in the following table 1. The parameters, which constitute the difference between them are:

- Channel type
- Gate oxide (GOX) thickness
- Operating voltage
- Drain extension

The table also contains a designation, which is chosen to be easy to memorize and distinguish between them.

Regarding the GOX types, definite figures can not be mentioned during this thesis, in order to protect the intellectual property of Infineon Technologies, but in general there are three different types in use. They differ only in thickness, which is normalized as a thickness unit (tu):

- GOX1 is 4.42 tu thick
- GOX2 is 0.42 tu thick
- GOX3 is 1 tu thick

One tu is a value, normalized to the thickness of GOX3.

The voltage ranges imply which of the devices are the high voltage power MOSFETs and which are the lower voltage lateral MOSFETs. Finally some of the devices have a drain extension and some do not. The drain extension is a shallow trench isolation (STI) between gate and drain, so that the conducting channel gets extended and the source-drain voltage can be increased. In

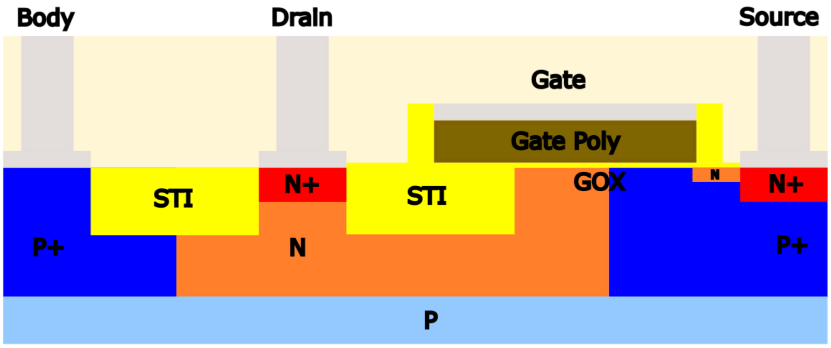


Figure 2: Schematic crosssection of an NDMOS with STI between gate and drain as drain extension.

figure 2 a schematic cross section of one of the NDMOS transistors with drain extension can be seen.

Name	Channel Type	GOX type	Operating Voltage / (V)	Drain Extension
NCAP	n-channel	1	5	no
NMOS-7V	n-channel	1	7	no
NMOS-5V	n-channel	1	5	no
NMOS-1.5V	n-channel	2	1.5	yes
PDMOS-60V	p-channel	3	60	yes
NDMOS-60V	n-channel	3	60	yes
NDMOS-40V	n-channel	3	40	yes

Table 1: Table of the seven new semiconductor devices on the BIRD, including information about their channel type, gate oxide type, operating voltage given in volts (V) and drain extension.

As can be concluded from table 1, the PDMOS-60V is a DMOS devices that operates in a high voltage range of 60 V. It has a drain extension and the 1 tu GOX3. It is also the only p-channel device on the chip.

By using the information in table 1 accordingly for the remaining devices one can easily understand their properties and purpose.

Despite the fact that some particular details about the new devices are not discussed in this thesis, the underlying physical principles remain unchanged. The following chapter 3 will therefore provide all the necessary information to understand the device's functionalities. This is done by first examining a MOS capacitor, followed by regular lateral MOSFETs with an addition about drain extensions in MOSFETs and power DMOS.

3 Metal oxide semiconductor field effect transistors (MOSFETs)

The underlying physics and equations, governing MOSFETs are of great importance, when it comes to understanding the testing of the functionality of single devices. Especially, when evaluating test data, measured at different conditions, it is necessary to understand what influences the environmental conditions, like e.g. temperature, have. When a device fails a test, it is helpful to have the equation, describing the device that failed the test at hand. Therefore, not only will the physics of the devices be explained in this chapter, but also the equations of the most relevant MOSFET parameters will be derived. The natural first step, of understanding the physics of MOSFETs, is to analyse the MOS capacitor, because the gate contact in a MOSFET basically is nothing other than that.

3.1 MOS capacitor

In one sentence, a MOS capacitor is a metal-oxide-semiconductor device, which purpose it is to store charge, like a conventional capacitor, but with a variable, voltage-controlled capacitance.

In its simplest form, a MOS capacitor is depicted in figure 3. Between the metal contact on top and the semiconductor, is an insulating oxide layer. At the bottom of the semiconductor, there is also an ohmic contact to ground.

In the following pages, only p-doped MOS capacitors are considered, since the one MOS capacitor on the BIRD is also a p-doped one and also most of the transistors are p-doped NMOS devices. Furthermore, the discussion will focus around what the working principle of such a device is. A detailed derivation of the capacitance curve will not be done here, since it would be too long and the BIRD mostly focuses on MOSFETs instead of MOS capacitors. A full description can be found in [3], which also strongly influenced the following derivations, formulas and figures.

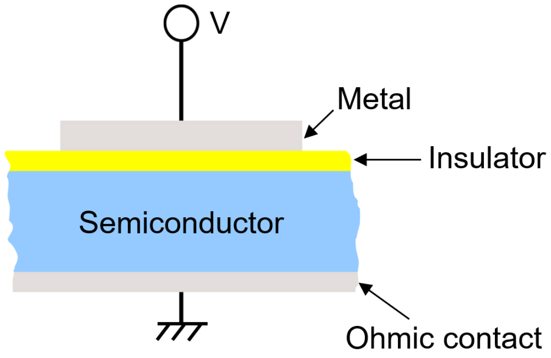


Figure 3: Crosssection of a p-doped MOS capacitor in its simplest form.

3.1.1 Ideal MOS capacitor

A MOS capacitor has to fulfill two criteria in order to be considered ideal:

(1) There can only be charges in the semiconductor and in the metal. This has to be true for all bias voltages. In other words, there are no oxide charges and charges in interface traps in the insulator.
(2) There is no charge carrier transport through the insulator.

The band diagram for an unbiased, ideal MOS capacitor looks like in figure 4, with ϕ_m and ϕ_p being the work function of the metal and the Fermi potential with respect to the valence band, respectively. χ is the electron affinity of the semiconductor and E_g the band gap energy. Furthermore, E_C, E_i, E_F, E_V are the conduction band energy, the Fermi energy for an intrinsic semiconductor, the Fermi energy and the valence band energy, respectively.

It is furthermore assumed, that the work functions of the metal and the semiconductor are the same. That means that the potential difference between the Fermi potential of the metal and the vacuum level and between the Fermi potential of the semiconductor and the vacuum level are the same. In real devices this does not necessarily have to be the case.

Described with the potentials shown in figure 4, this means that:

$$\phi_{ms} = \phi_m - (\chi + \frac{E_g}{q} - \phi_p) = 0 \tag{1}$$

This also implicates that flat-band condition is reached, when there is no bias

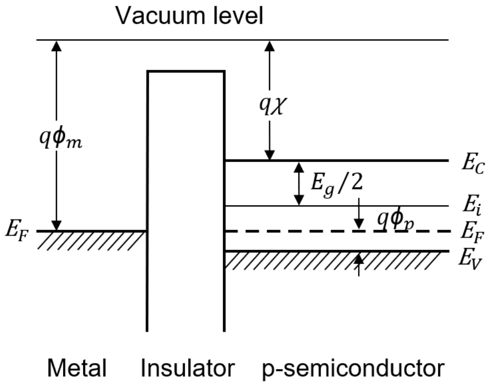


Figure 4: Band diagram of a p-doped ideal MOS Capacitor, with no bias voltage.

voltage applied to the device. Upon biasing the gate, there are several states the MOS capacitor can be in, which are related to the surface space-charge at the semiconductor-insulator interface.

The surface space-charge density is given in units of (C/cm^2) and is the amount of charge per area at the surface of the semiconductor, at the interface with the insulator. It is dependant on the surface potential as well as the electric field. The working principle of a MOS capacitor can be shown by their relations.

The intrinsic (or midgap) potential $E_i(x)/q$, with respect to the semiconductor potential at the bulk $E_i(\infty)/q$ is defined as $\Psi_p(x)$:

$$\Psi_p(x) = \frac{E_i(x) - E_i(\infty)}{q} \tag{2}$$

As can be seen in figure 5, $\Psi_p(0) \equiv \Psi_S$, where Ψ_S is the surface voltage. The potential difference between the Fermi potential and the midgap is Ψ_{Bp}. The mode in which the device is in, depends on whether Ψ_S is positive or negative and if it is greater or smaller than Ψ_{Bp}, which is dependent on the bias voltage. In figure 5 the p-type semiconductor is biased with a positive voltage, which causes the bands to bend downwards near the interface and Ψ_S to be larger than 0.

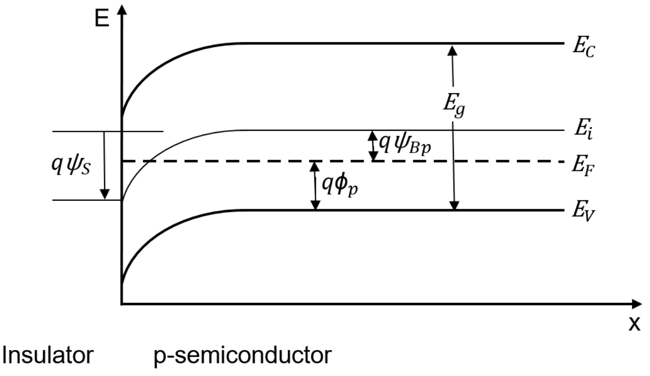


Figure 5: Band diagram of a p-type semiconductor. The downward bending bands are due to a positive bias voltage $V > 0$.

The relation between the minority carrier density (electron density), the majority carrier density (hole density) on the semiconductor surface and the voltage $\Psi_p(x)$ is given by:

$$n_p(x) = n_{p0}\ e^{\beta\Psi_p(x)} \tag{3}$$

$$p_p(x) = p_{p0}\ e^{-\beta\Psi_p(x)} \tag{4}$$

where $\beta = q/k_BT$ and n_{p0} and p_{p0} are the minority and majority carrier densities at equilibrium in the bulk of the semiconductor. At the surface of the semiconductor ($x = 0$), $\Psi_p(x)$ in the equations above, becomes Ψ_S.

So the carrier concentration at the semiconductor-insulator interface depends on the surface potential and therefore on the bias voltage applied to the metal on top of the MOS capacitor.

The following relations, which correspond to the different states the MOS capacitor is in, can now be derived:

State	**Surface voltage**	**Gate voltage**	**Electron density**
Accumulation	$\Psi_S < 0$	$V < 0$	$n_p(0) < n_{p0}$
Flat-band	$\Psi_S = 0$	$V = 0$	$n_p(0) = n_{p0}$
Depletion	$\Psi_{Bp} > \Psi_S > 0$	$V > 0$	$n_p(0) > n_{p0}$
E_F at midgap	$\Psi_{Bp} = \Psi_S$	$V > 0$	$n_p(0) > n_{p0}$
Weak inversion	$2\Psi_{Bp} > \Psi_S > \Psi_{Bp}$	$V > 0$	$n_p(0) > p_p(0)$
Strong inversion	$\Psi_S > 2\Psi_{Bp}$	$V > 0$	$n_p(0) > p_{p0} > p_p(0)$

Table 2: States of the MOS capacitor, with their respective surface potential relations, gate voltage relations and electron densities.

The three key states the device can be in are (1) accumulation, (2) depletion and (3) strong inversion. Those three states will be described in more detail, because they are the key points of how a MOS capacitor and a MOSFET work. In figure 6 the band diagrams for (1), (2) and (3) are depicted for a p-type semiconductor.

Accumulation: When biasing the metal on top of the device with a negative voltage, the potential causes the bands of the semiconductor to bend. For a negative voltage the bands bend upwards, which, for a p-type semiconductor, means that the number of majority carriers (holes) at the interface to the insulator increases. This is because the charge carrier density in the valence band depends on the energy difference ($E_F - E_V$) between the Fermi energy and the valence band. Since the Fermi energy does not bend, an accumulation of holes is happening. Due to this, charge is stored near the interface and the capacitance of the device increases.

Depletion: During depletion, the metal is biased with a positive voltage. Since the semiconductor is assumed to be in flat-band condition when there is no bias voltage, the bands start to bend downwards after only applying a small positive voltage. Because the energy difference between the Fermi energy and the valence band is now increasing, the number of holes is decreasing. At the same time, the voltage is not large enough to attract many minority carriers (electrons) yet, therefore the region near the interface lacks of charge carriers and the capacitance is therefore decreasing.

Strong inversion: If the positive voltage is surpassing Ψ_{Bp}, the holes are not only being pushed from the interface, but electrons are also being attracted. If the positive voltage is large enough that the number of electrons exceeds the number of holes near the interface, the device is in inversion. The device is in strong inversion when the surface potential is twice as high as Ψ_{Bp}. Then, the number of electrons near the interface is larger than the number of holes in the bulk of the semiconductor. Because of this increase in charge carriers near the interface, the capacitance of the device rises again, this time, however, due to the minority carriers.

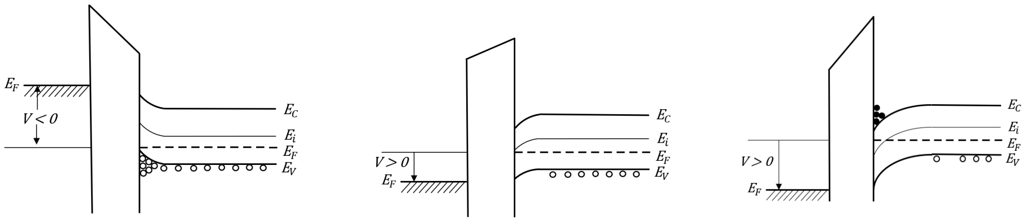


Figure 6: Band diagrams for an ideal MOS capacitor during accumulation (left), depletion (center) and inversion (right).

As mentioned above, all of this is linked to the surface space-charge density Q_S. Finding a solution for Q_S as a function of Ψ_S is crucial for deriving the capacitance curve of a MOS capacitor. Since the focus of this thesis is not on MOS capacitors, the long derivation for Q_S is only referred to and not done here.

So in order to get a complete and thorough derivation of the surface space-charge density and the capacitance curve of the ideal MOS capacitor, see [3].

3.1.2 Real MOS capacitor

As the previous section was only limited to ideal MOS capacitors, it is useful to mention what the main differences are to a real MOS capacitor. This is also quite useful, since it also foreshadows some of the failures that can arise in real devices.

As already mentioned at the beginning of the last section, there are two criteria that an ideal MOS capacitor has to fulfill. Real devices, however, can have charges in the oxide and in interface traps. This enhances charge transport through the insulator.

Not only do these charges influence the capacitance curve of the device and therefore, also its performance, but they also enhance an early breakdown of the insulator material. This is a most unwanted behaviour, because the charges can then travel freely from the metal through the insulator into the semiconductor and vice versa.

This mechanism is called a failure mechanism and are described in detail in chapter 4. However, the mechanisms described there are mainly regarding MOSFETs. The following pages therefore, are giving an overview of the physics behind MOSFETs and what parameters are of interest when characterizing them. This is followed by what types of MOSFETs exist and are then focusing on the type that is used on the BIRD chip, which is described in more detail.

3.2 Regular lateral MOSFET

Like already mentioned, a MOS capacitor is part of the MOSFET. A cross section of a MOSFET in its simplest form is given in figure 7. More precisely, the gate contact of a MOSFET can be view as one. Left and right to the gate contact are the source and drain contacts, which are on top of an n-doped region (when assuming a p-doped bulk). This makes a MOSFET a four teminal device (source, gate, drain and bulk). However, often times the bulk and the source are connected to each other and held at the same potential.

Upon biasing the gate with a large enough positive voltage, inversion beneath the gate contact is reached, which connects the n-doped regions beneath source and drain. Usually they are separated by the p-doped bulk. The inversion layer forms a conducting channel through which a current flow from source to drain is possible. When the drain contact is biased with a positive voltage as well, the electrons start to drift from source to drain, establishing said current flow.

The basic working principle of a MOSFET in its most basic form can be seen in figure 7. On the left hand side the gate of the device is unbiased and on the right hand side a conduction channel has been formed, because the gate is biased with a positive voltage.

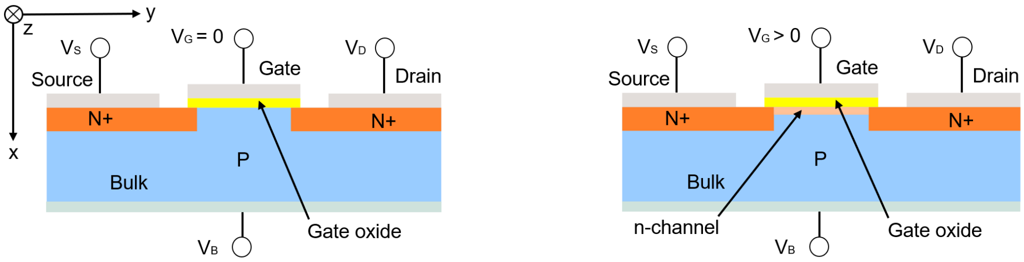


Figure 7: Cross section of a basic MOSFET with zero gate bias on the left and positive gate bias, together with an inversion channel on the right.

Devices that have a conduction channel which is formed by electrons and have a p-doped bulk, are called NMOS. On the other hand, when the conduction channel is formed by holes in an n-doped bulk, it is called a PMOS. When the gate is unbiased, there can be two cases at hand. One, where the device is conducting with zero gate bias (normally-on), called depletion-mode and one, where the device is not conducting with zero gate bias (normally-off), which is called enhancement-mode. Those two modes exist for both the NMOS and PMOS. Since the majority of devices on the BIRD are NMOS transistors in enhancement-mode, the following pages will also focus on those.

Another noteworthy distinction between lateral MOSFETs is that the conducting channel can be at the surface, at the interface to the insulator or buried further down in the bulk. The latter one is called buried channel device and has its own characteristics. The MOSFETs on the BIRD, however, are surface inversion channel devices, so the buried channel devices will not be discussed here. Further readings on them can be found in [3].

Threshold voltage: One important parameter that was not mentioned during the discussion of the MOS capacitor, is the threshold voltage V_T. This is the voltage at which the MOS capacitor has built up the inversion channel to such an extent, that the MOSFET can be considered as 'on'.

This voltage is reached, when the electron density in the inversion channel is as high as the hole density in the bulk of the device. It can be interpreted as the gate bias beyond the flat-band, just starting to induce an inversion charge sheet. Written in the potentials of the discussion about the MOS capacitor, the threshold voltage is:

$$V_T = V_{FB} + 2\Psi_{Bp} + \frac{\sqrt{2\epsilon_s q N_A 2\Psi_{Bp}}}{C_{ox}} \tag{5}$$

V_{FB} corresponds in this formula to the voltage that needs to be applied, in order to achieve flat-band condition. In the previous discussion about MOS capacitors, this voltage was zero, which was merely an assumption at the time. $2\Psi_{Bp}$ is the voltage drop across the semiconductor and the oxide layer at strong inversion. The square root term is the total depletion layer charge.

This is one of the most important quantities when characterizing a MOSFET, because it also depends on parameters like the acceptor doping N_A, the dielectric material and the oxide thickness. They are included through Ψ_{Bp} and C_{ox}.

Using the introduced threshold voltage, the next pages will focus on some of the device characteristics and important parameters of MOSFETs. To be precise, the MOSFETs which will be used for this are lateral surface inversion n-channel MOSFETs in enhancement-mode.

3.2.1 Current-voltage characteristics

One of the most obvious and underlying device characteristics is the current-voltage characteristic (I-V curve). The following derivation will follow [4] and hence only the most important steps and formulas will be written in this thesis.

Because a real MOSFET is a 3-dimensional device, with different charge distributions and electric fields in the x-,y- and z-direction, a complete description of the device can only be done through numerical simulations. Therefore, a one-dimensional model will be introduced, which will preserve the main aspects of the working principle.

The following assumptions are made:

1. An ideal MOS capacitor is used as the gate structure.
2. The drain current consists only of drift current.
3. The electron mobility μ_n and hence their velocity v_n is linear with the electric field, which is true for low electric fields.
4. The electric field in x-direction is much larger than the electric field y-direction (along the channel).

Assumption number (2) implies that there are no short or long channel effects at hand, like sub threshold voltage or channel length modulation.

Assumption number (3) is only valid when the electric field in the channel is below 3 kV/cm, which modern devices normally exceed.

Assumption number (4) is the gradual channel approximation, which can be applied when the gate length L is large compared to the channel depth.

The first step is to introduce the relation between the channel charge, as a function of y (the axis along the channel) and the gate voltage.

In inversion, the channel charge is:

$$Q_S = C_{ox}(V_{GS} - V_T - V_c(y)) \tag{6}$$

with V_{GS} being the voltage between the gate and the source contact and $V_c(y)$

being the voltage along the channel.

$$V_c(y) = 0 \qquad \text{at source } (y = 0) \tag{7}$$

$$V_c(y) = V_{DS} \qquad \text{at drain, } (y = L) \tag{8}$$

with V_{DS} being the voltage between source and drain and L being the channel length.

The current density for the one-dimensional gradual channel approximation is defined as:

$$j = \frac{I}{Z} = -Q_S v_d \tag{9}$$

where v_d is the drift velocity of the charge Q_S (electrons) and Z is the channel width. In order to establish a relation between the channel voltage and the current, one substitutes the drift velocity with:

$$v_d = -\mu_n \frac{dV_c(y)}{dy} \tag{10}$$

This results in an expression for the drain current, dependent on the channel voltage:

$$I_D = Q_S \mu_n \frac{dV_c(y)}{dy} Z \tag{11}$$

This is a first order differential equation. To solve it, Q_S needs to be substituted and one needs to integrate over the channel length (from source, $y = 0$ to drain, $y = L$).

Using $V_c(L) = V_{DS}$, the solution to this equation is:

$$I_D = \frac{\mu_n Z C_{ox}}{L} \left(V_{GS} - V_T - \frac{V_{DS}}{2} \right) V_{DS} \tag{12}$$

When increasing the drain voltage, it reaches a point, where the term in the brackets becomes zero. At this point, the formula for I_D and the used model is reaching its limits and another description has to be found for the drain current. The inversion channel is pinched off at the drain end. This point defines the saturation voltage $V_{DS}(sat)$:

$$V_{DS}(sat) = V_{GS} - V_T \tag{13}$$

Linear region: For small V_{DS} values, the quadratic term in equation 12 is negligible and the linear term dominates. This part of the drain current is called the linear regime. At higher voltages the quadratic term dominates, at which point pinch off is near. For the linear region of the drain current the following approximation can be used:

$$I_{D_{lin}} = \frac{\mu_n Z C_{ox}}{L}(V_{GS} - V_T)V_{DS} \tag{14}$$

Saturation region: When the channel gets pinched off by the increasing drain bias, the quadratic term starts taking effect and the current curve enters the saturation regime. The drain current in saturation is not increasing anymore and can be calculated by inserting $V_{DS}(sat)$ into equation 12:

$$I_{D_{sat}} = \frac{\mu_n Z C_{ox}}{2L}(V_{GS} - V_T)^2 \tag{15}$$

As can be seen, the drain current is now not dependent on the drain voltage, but only on the gate voltage.

In figure 8 the drain current curve can be seen for several different gate voltages. Once again, this is the I-V-characteristic of a MOSFET, which fulfills all the previous assumptions.

The lateral MOSFET, which was described in this chapter, is a good way to start deriving the most important relations and working principles of such devices. It is, however, by far not the only MOSFET structure, which is in application in modern electronics. As already mentioned earlier, even this simple lateral MOSFET can be an NMOS or PMOS transistor, which can either be in enhancement or saturation mode.

The following pages will therefore give an overview over some of the most important types of MOSFET.

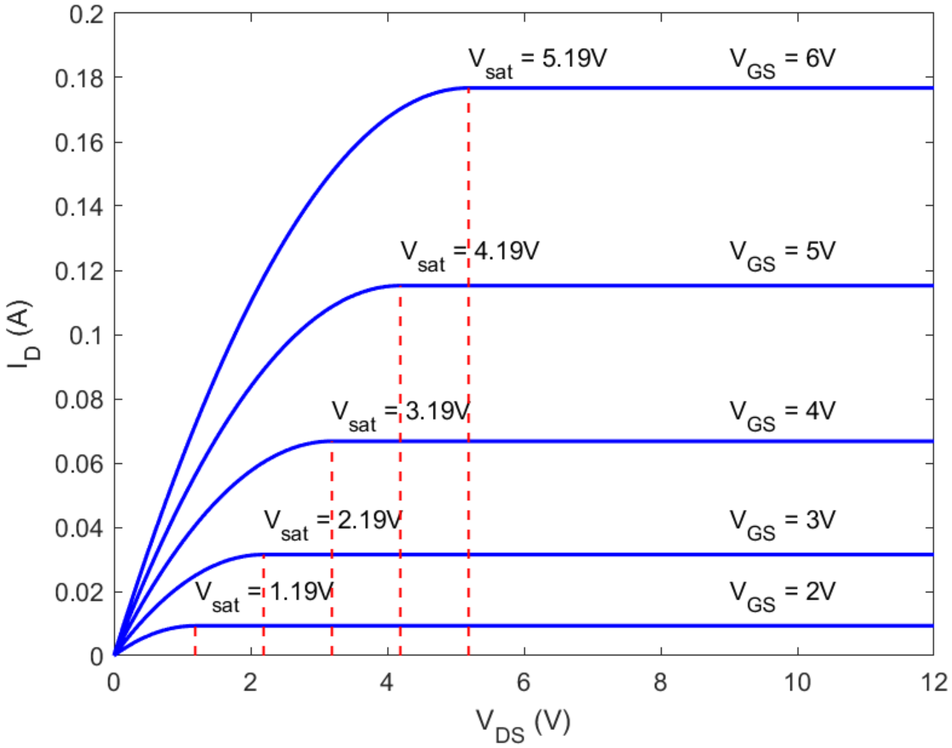


Figure 8: Drain current I_D as a function of the drain to source voltage V_{DS} for five different gate voltages V_{GS}. Clearly visible is the linear range at the beginning and the saturation region, where the drain current becomes constant.

3.3 Types of MOSFETs

When speaking of the conventional lateral MOSFET, the various different types have already been mentioned in chapter 3.2. Of course there are many ways to combine the different ones. One way is, to build NMOS and PMOS next to each other. This is called complementary MOSFET (CMOS).

CMOS: In CMOS technology, the PMOS is on top of an n-doped well, which itself, is inside the p-doped substrate. When this is the primary used block in the design, one speaks of CMOS architecture. A cross section of one basic CMOS can be seen in figure 9. CMOS architecture is used in many applications, especially in digital applications such as memory and logic. One of its major advantages is the low energy consumption and dissipation. For further reading on CMOS technology see [5].

The digital part of the BIRD is built with a CMOS architecture too. Furthermore, all the inverters on the enable line of the devices on the BIRD, are just basic CMOS inverters.

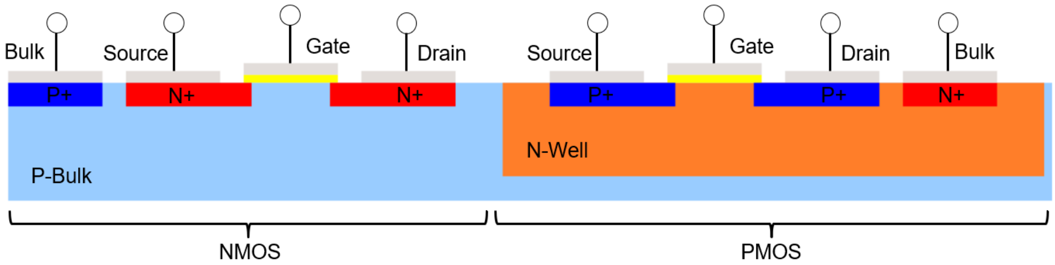


Figure 9: Cross section of a basic CMOS with a p-doped substrate.

Power MOSFET: If the goal is to switch high voltages (up to kV) and carrying more current (beginning with 1A), CMOS and regular lateral MOSFETs are limited, because of either gate oxide breakdowns, source/drain-body breakdowns or punch through of the two junctions between source and drain. For high power applications, power MOSFETs are used. Some of the MOSFETs on the BIRD are also power MOSFETs, ranging up to 60V. This is more than what is typically used in CMOS technology, although, it is still at the lower voltage side of the power MOSFET spectrum.

There are several different ways of how a power MOSFET can be realized. For example there are double diffused MOSFETs (DMOS), which can either have a lateral or a vertical structure. Furthermore, there are v-groved MOSFET (VMOS) and u-shaped notch MOSFET (UMOS), both of which have a trench structure. Cross sections of all of these can be found in figure 10.

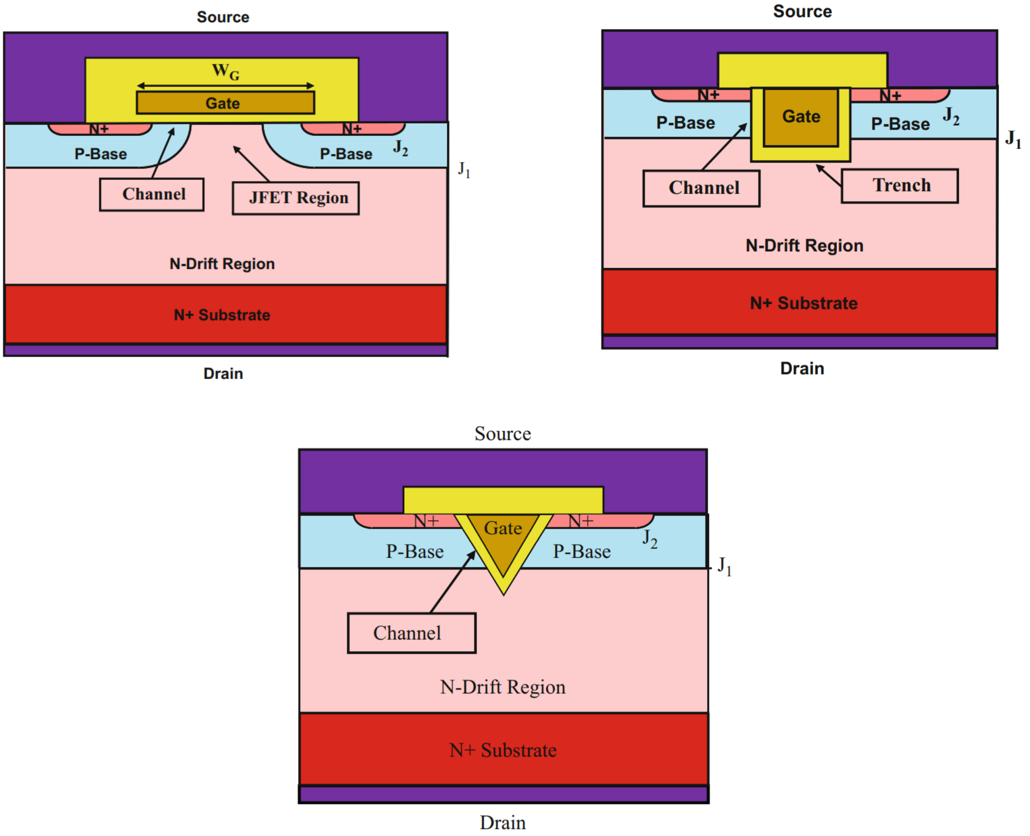


Figure 10: Cross section of a DMOS on the top left, a VMOS in the bottom middle and a UMOS on the top right.[6]

The reason as to why these devices are more resilient against high voltages, and currents is, that their geometry, together with their doping profile allows such high power. In chapter 3.4, the DMOS structure and its working principles will be explained in more detail.

FinFET: One other big family of MOSFET are called FinFETs, which are non-planar devices. They arose out of the need of scaling the devices down, in order to build more of them on the same area. One way to do this, is by either decreasing the channel length or the oxide thickness. This method, however, only works up to some limit, at which quantum effects (like tunnel current through the gate oxide) start interfering with the performance of the devices. As a consequence, a new architecture of MOSFET was thought of, which is the FinFET. In figure 11, a basic FinFET structure is depicted. The gate is here covering the channel (the fin) on three sides. Some of the advantages of FinFETs include a better control over the channel, suppressed short-channel effects and a faster switching speed. The leakage current problem, caused by the mentioned tunneling is solved as well. For further reading on FinFETs, see [7].

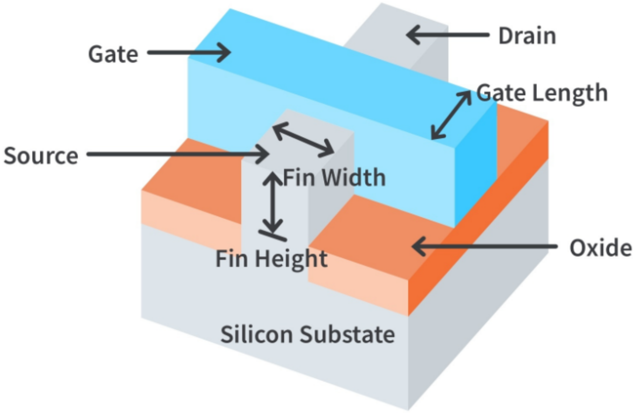


Figure 11: Basic structure of a FinFET.[8]

3.4 Double diffused power MOSFET (DMOS)

As already briefly mentioned in chapter 3.3, the DMOS, which this chapter focuses on, can be realized in two ways. Once as a lateral DMOS (LDMOS) and once as a vertical DMOS (VDMOS). A cross section of a VDMOS can be seen in figure 10 on the top left side. A cross section of an LDMOS can be seen in figure 12. Since the power MOSFETs on the BIRD are all LDMOS, the following pages will only focus on them.

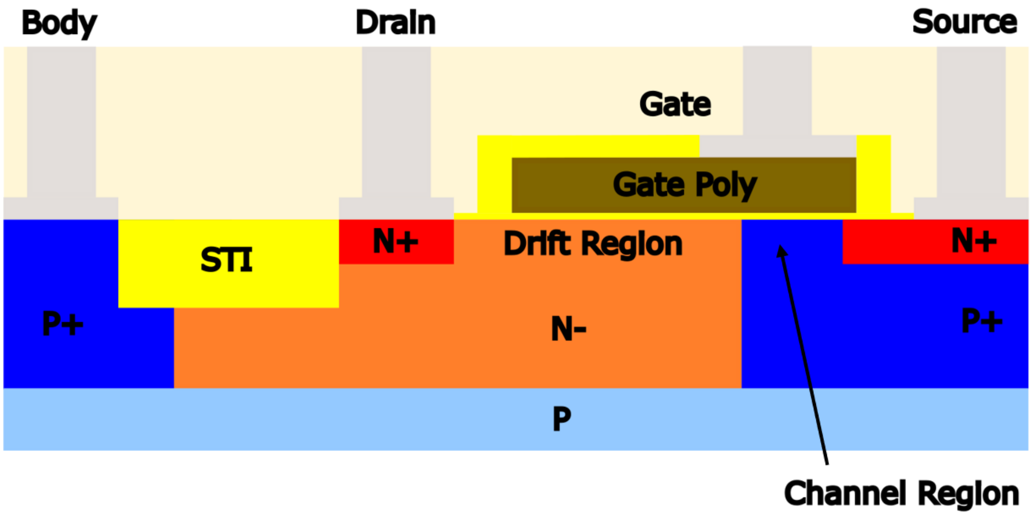


Figure 12: Cross section of an LDMOS.

In contrast to figure 7 (the simple MOSFET), the body contact is not at the bottom, but on top of the MOSFET. Furthermore, it is separated from the drain contact via an STI, in order to prevent a junction breakdown when biasing the drain contact. The main difference, however, lies within the two different dopings beneath the gate. While a regular NMOS has only a p-doped region under the gate oxide, the LDMOS has an n-doped region as well.

This lightly n-doped region is called the drift region, while the p-doped one is called the channel region. Since the inversion channel will be created in the p-doped channel region, the device in figure 12 is an NMOS as well.

The drift region is the reason why LDMOSs can be operated with greater voltages than regular MOSFETs. When solving Poisson's equation for an abrupt junction, like between the channel region and the drift region, the resulting electric field has a triangular shape. The slope of this electric field is determined by the doping concentration, where a higher doping concentration results in a steeper slope. Furthermore, the depletion width is also depending on the doping concentration, as well as the maximum of the electric field. Whether the transistor can withstand high voltages or not. depends on the doping concentration and the semiconductor material. A lower doping concentration results in a higher breakdown voltage, as well as in a wider depletion width, as can be seen in equation 16 and 17.

$$V_B = \frac{\epsilon_S E_C^2}{2qN_D} \tag{16}$$

$$W_D = \frac{2V_B}{E_C} \tag{17}$$

Where V_B is the breakdown voltage of the reverse biased junction, E_C the critical electric field, ϵ_S the dielectric constant of the semiconductor, N_D the donor concentration in the n-doped drift region and W_D the depletion width of the drift region. [6]

3.4.1 Device characteristics

At the end of chapter 3.2, the current-voltage characteristics of a simple MOSFET has been derived. This derivation, however, was made under a few assumptions, which do not necessarily apply for real devices.

In the following pages, the most important effects and characteristics, which are used and observed in real LDMOS devices are described.

On resistance: The on resistance (R_{DSon}) is one of the most important device parameters, because it governs the power dissipation. It is the resistance between source and drain, when the device is considered ’on’.

The resistance of a semiconductor is dependent on its doping concentration. With increasing doping concentration, the resistance gets lower. This is valid for p- as well as for n-doping, the resistivity for most of the n-doped semiconductors, however, is lower than that of the p-doped ones, since the intrinsic electron mobility is higher than the hole mobility. The ideal specific on-resistance for an n-doped silicon semiconductor is around three times as low as for a p-doped one, which is the reason why most of the power MOSFET devices are NMOS. [6]

As there are several different doping regions from drain to source, the total resistivity of the LDMOS device is a sum of several separate resistances, one for each doping region, as well as the contact resistances of the source and drain contact.

$$R_{DSon} = R_{source} + R_{N+} + R_{ch} + R_D + R_{N+} + R_{drain} \tag{18}$$

R_{ch} is hereby the resistance of the channel region of the LDMOS, while R_D is the resistance of the drift region. R_{source} and R_{drain} are the contact resistances of the source and drain, while R_{N+} are the resistances of the highly doped n-regions beneath source and drain.

The channel resistance R_{ch} can be calculated by using the approximation for the drain current in the linear region as shown in equation 14, since most of the LDMOS devices are operated in the linear region, due to the lower channel resistance. [9]

$$R_{ch} = \frac{V_{DS}}{I_{D_{lin}}} = \frac{L_{ch}}{Z\mu_{n,ch}C_{ox}(V_{GS} - V_T)} \tag{19}$$

The drift resistance R_D additionally depends on the doping concentration of the drift region and the depth of the drift region d. It is then given by:

$$R_D = \frac{L_D}{q\mu_{n,D}N_D Z d} \tag{20}$$

Since these two terms dominate over the other contributors, R_{N+}, R_{source} and

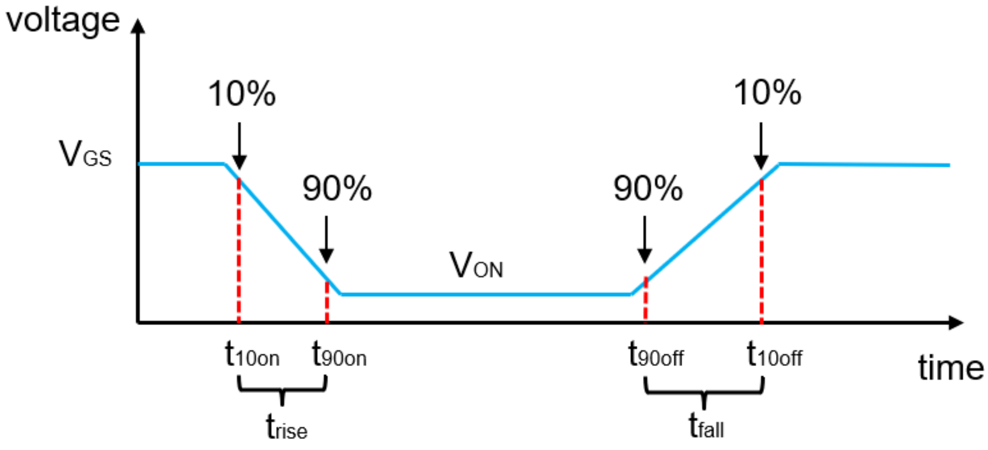


Figure 13: Schematic function of the fall (t_{fall}) and rise (t_{rise}) time of a MOSFET.

R_{drain}, the explicit formula for them are not included in this discussion, they can, however, be found in [6].

The on resistance is, furthermore, also a function of temperature, since quantities like the threshold voltage and the charge carrier mobility depend on the temperature.[6]

Fall / rise time: The fall time is defined as the time that is necessary, for the voltage between source and drain while 'off', to drop by 10 % of itself, to 90,% of itself. This happens after the gate voltage is increased and the device turns on.

Similarly, is the rise time defined as the time the device needs, for its voltage between source and drain to increase by 10 % of itself to 90 % again. This happens after the gate voltage has decreased and the device starts to switch off.

Figure 13 shows the change of the voltage between source and drain, as a function of time, together with the associated fall and rise times. As shown in the figure, the fall and the rise time do not have to be the same, however, minimizing both is wanted in the industry. This is due to on one hand the faster switching speed allows for higher switching frequencies and less power dissipation.

Power dissipation: The power dissipation is the reason why the switching speed and the on resistance are of high importance when characterizing power MOSFETs. High power dissipation heats up the transistor and influences its performance.

There are two parts that sum up for the total power dissipation:

$$PD_{tot} = PD_{resistive} + PD_{switching} \tag{21}$$

where $PD_{resistive}$ is the power dissipation due to the on resistance and $PD_{switching}$ is the power dissipation due to the switching performance. [10]

The power dissipation resulting from the on resistance is calculated through using the equation for power combined with Ohm's law:

$$PD_{resistive} = I_D^2 * R_{DSon} \tag{22}$$

This is the power that is dissipated in the device during on-state operation, which increases linearly with the on resistance, which makes a lower on resistance desirable. The power dissipation resulting from the switching performances is only taking effect during the brief moments of switching the device on/off.

$$PD_{switching} = PD_{on} + PD_{off} = \frac{t_{rise}}{2T} I_D V_{DS} + \frac{t_{fall}}{2T} I_D V_{DS} \tag{23}$$

Where T is the switching period, which is the time between switching the device and the time where it is off again. That means, that a short fall and rise time, together with a long switching period, minimizes the power dissipation. [9]

3.4.2 LDMOS optimization

In order to improve some of the effects and characteristics mentioned above, the geometry of the device can be altered. Beginning with the doping concentration, the diffusion depth, the channel length and many more. Some of the most important optimization strategies are discussed in the following pages. Since there are of course many different ones, a focus is laid on the ones that have also been used on the transistors on the BIRD.

RESURF method: One aspect that needs to be considered in the design of an LDMOS, is that there is a trade-off between the breakdown voltage of the body-drift diode (BV_{DSS}) and the on resistance. The breakdown voltage of the body-drift diode, is the voltage that needs to be applied, so the junction between the channel region (which is part of the bulk) and the drift region breaks down. After that, a current starts flowing between source and drain, even though no gate voltage is applied.

When designing devices that need to have a large BV_{DSS}, one way to do so is to make the drift region longer and more lightly doped. This, however, increases the on resistance of the device and makes the devices larger, both of which is unwanted. This can be understood, by looking at equation 20.

An alternative way to increase BV_{DSS}, however, is to use the reduced surface field (RESURF) method, which distributes the space charge region over the whole drift region. A consequence of this is, that the potential drop occurs not only at the junction anymore and the maximum electric field gets reduced. This is realised, by reducing the depth of the drift region. [11] [12]

Using this method, both the doping concentration in the drift region can be increased and the length can be decreased, which decreases the on resistance and therefore the energy dissipation. Decreasing the depth of the drift region, increases the on resistance as well, this effect, however, gets canceled out be the new doping concentration and channel length.

Gate contact extension: One other alteration of the classical LDMOS, like fig. 12, that decreases the on resistance as well, is extending the gate contact further over the drift region. This way, accumulation is achieved at the interface between the drift region and the gate oxide, when the inversion channel is formed. This increases the free charge carriers in the drift region and introduces a new resistance R_A, which contributes to the total on resistance. R_A is smaller than R_D and reduces the overall resistance, because some percentage of R_D's contribution to $R_{DS_{ON}}$ gets substituted by R_A. [13]

Drain extension: The last optimization that is going to be introduced in this chapter is the drain extension. As already briefly mentioned in chapter 2.2, the drain extension is an STI between the gate and the drain contacts. So inside the drift region, there is isolating SiO_2 which serves the purpose to extend the drift path of the electrons and thus increasing the maximum voltage before breakdown. [14] [15]

Applying all those optimizations to the LDMOS depicted in figure 12, the new cross section looks like in figure 14.

Of course, the six different MOSFETs on the BIRD do not all have the same cross section and the cross section in figure 14 is a still slightly simplified one, in order to show the most important features of the actual devices.

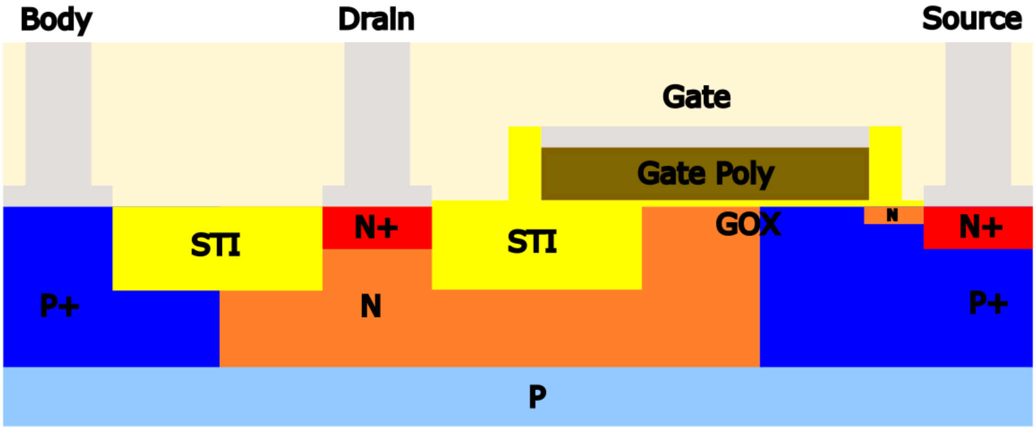


Figure 14: Cross section of an LDMOS with a gate contact extended into the drift region, a drain extension and increased drift region doping, due to the RESURF method.

Closing with this cross section, the discussion about MOSFETs, LDMOS and what their working principles are, is finished. However, since this thesis is not only focusing on how these devices work, but also investigates potential problems, the following chapter is about failure mechanisms.

It starts with industry standards and to which quantities failures can be tolerated, which results in an discussion about the early life failure rate. Afterwards, some specific failure mechanisms and their consequences will be discussed in detail.

4 ELFR and Failure mechanisms

After having established the proper working principles for the power devices on the BIRD, the next step is to describe the most likely mechanisms that can cause them to fail. As already hinted in chapter 3.1.2, failures occur in real devices and have been disregarded in the previous discussions about idealized devices. Before getting to the failures themselves, however, the next few pages will go deeper into the consequences a high rate of failing devices has. It starts with what the industry standards are and will then move towards a discussion about how to mathematically describe the rate by which devices fail and how to lower this rate in order to fulfill the introduced industry standards.

4.1 Industry standards

There are several norms and standards, which should be complied, when manufacturing semiconductor devices. Especially in the automotive industry, where the devices on the BIRD will be used in, assuring reliability is crucial, since malfunctioning devices might, in the extreme case, lead to heavy injuries or even death.

It is essential for the devices and products that are sold, to comply with the above mentioned industry reliability standards, which implies which devices that fail at an early point in their life need to get sorted out, before being sold.

The standard, which is used at Infineon Technologies (and has been adopted by most European manufacturers), is the ISO 26262, which is a sector specific adaptation of the functional safety standard IEC 61508 for automotive electric or electronic systems. It is a risk based safety standard, where hazardous situations are assessed and safety measures are defined to avoid and control failures. Its main goal is to reduce harm to humans and provide an automotive specific risk based approach to minimize the effects of systematic and random hardware failures of electric or electronic automotive systems.

One of the key aspects of the ISO 26262 is the automotive safety integrity level (ASIL) risk level ranking for technical processes, going from ASIL A (highest allowable risk) to ASIL D (lowest allowable risk). This system ranks various requirements of components by the levels of allowed risk. The driving factors, which determine risk, in the way the ISO 26262 defines it, are:

- Frequency of occurrence (F)
- Driver controllability (C)
- Potential severity (S)

Controllability and potential severity are again divided into levels, which rank from C0 (controllable in general) to C3 (difficult to control) and from S0 (no injuries) to S3 (Life-threatening injuries) respectively. While controllability corresponds to the likelihood that a driver can act to prevent hazard, severity is an estimate of the harm that can occur in a hazardous situation. The frequency of occurrence is furthermore determined by the failure rate (λ) times the exposure (E).

$$F = \lambda * E$$

Exposure is the probability of a human's exposure to a hazard in terms of frequency or duration. It gets classified into 5 levels E0 (lowest exposure) to E5 (highest exposure).

The failure rate λ is the only thing that can be controlled by engineers and is a property of the system/component/device. In the context of ASIL ratings, it is defined as the allowable failure rate, with ASIL D having the lowest allowed failure rate. Because the failure rate, used in the ISO 26262 mainly focuses on systems and components, like a battery management chip for electric cars or the sensor which activates the airbag, it is not directly applicable for the fabrication process.

Due to these safety standards, provided by the ISO 26262 norm, it is important for the manufacturing process to be adjusted in a way that allows it to comply. This starts not only by manufacturing the products, which are sold to the customer, in a way that they fulfill the standards, but by manufacturing every component in every integrated circuit in the product, such that this can be achieved. Therefore, also the semiconductor devices in the ICs, have to have a low enough failure rate. But what is the failure rate and how can it be lowered, if it is too high in order to comply to the ISO 26262?

4.2 Early life failure rate (ELFR) and burn-in

When manufacturing a large amount of the same product with the same process, chances are that some of them are better or worse in certain ways than others. This is a very general way of expressing the fact, that during most of the steps of the manufacturing process small deviations from the ideal exist, which results in some of the products or, in the case of the BIRD, some of the devices on the chip, to have a better or worse performance than others.

During the lifetime of any large quantity of manufactured device, there are several phases the product goes through. Because of the mentioned deviations from the ideal manufacturing process, some of the devices do not work properly from the start, while others tend to fail later in their life. What all these devices have in common, is that they fail to work eventually. How long it takes for them to fail is called the time-to-failure (TF). Naturally, the longer this time is, the better the reliability of the device.

At this point it needs to be said, that the following discussion about the time-to-failure, degradation and failure rates is strongly influenced by [16] and presented in a condensed version with only the information relevant to the BIRD burn-in study in it. In [16] the mathematical derivations are explained in more detail and some concepts which will be merely mentioned here, are discussed more thoroughly.

4.2.1 Time-to-failure and degradation

The time-to-failure is reached, when one of the device's parameters has shifted by such an amount, that functionality can no longer be guaranteed. In order to estimate when this time has been reached, it is necessary to model how the parameters of the device degrade with time. Assuming the device parameter S under investigation changes monotonically and relatively slowy (this is true for the MOSFET parameters introduced in chapter 4), a Taylor expansion around $t = 0$ can be made:

$$S(t) = S_{t=0} + \left(\frac{\partial S}{\partial t}\right)_{t=0} t + \frac{1}{2}\left(\frac{\partial^2 S}{\partial t^2}\right)_{t=0} t^2 + ... \tag{24}$$

By approximating the higher order terms with a power-law exponent m, the series in equation 24 can be written as:

$$S = S_0(1 \pm A_0 * t^m) \tag{25}$$

with A_0 being a device dependent coefficient. Both parameters, m and A_0 are variables which can be determined by evaluating existing empirical degradation data. Whether A_0 is positive or negative, will lead to a decrease or an increase of S, respectively. Decrease and increase can both lead to a fail.

Two other time dependent degradation models which are in use, are the exponential and the logarithmic degradation models. However, they are more rarely used than the power-law. Closer information on them can be found in [16].

From equation 25 the time-to-failure can be calculated in terms of the degradation, where it is necessary to define the maximum allowed degradation S.

$$t = TF = \left(\frac{1}{\pm A_0} \left(\frac{S - S_0}{S_0} \right)_{crit} \right)^{\frac{1}{m}} \tag{26}$$

When m is equal to zero, which means no degradation at all, TF diverges to infinity. In figure 15 a decrease and an increase of an undefined parameter, as a result of degradation is depicted, together with the maximum allowed degradation for both decrease (black line at 0.5) and increase (black line at 2). The time-to-failure is indicated in red at the point where the degradation crosses the maximum or minimum allowed value, respectively. It can be seen that for higher positive values of A_0, the degradation reaches its critical value faster and the time-to-failure is shorter than for lower values. The same is true for lower negative values.

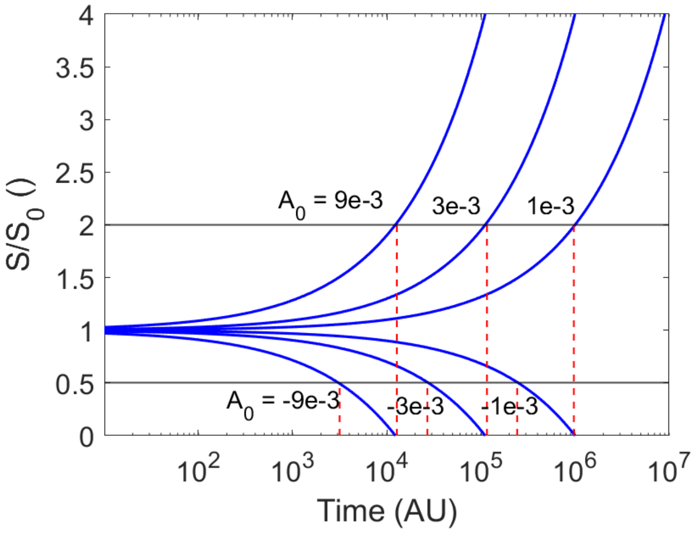


Figure 15: Degradation of an undefined parameter with respect to time, with several different values for A_0 and $m = 1$.

Even though A_0 is a device parameter, which depends on the material/ microstructure, it is not static itself. Amongst others, it can depend on the mechanical stress, temperature, the applied voltage, the chemical environment and every combination of them. In the case for the devices on the BIRD chip, the relevant dependencies, however, are the dependence on voltage and temperature. Therefore A_0 becomes the parameter which gets affected by the increase in temperature during burn-in and by the stressing of half of the devices with overvoltage:

$$A_0 = A_0(V, T) \tag{27}$$

Because of the slight difference of the manufacturing process of each device, that was mentioned at the beginning of this chapter, the time-to-failure is not the same for every device. TF itself is distributed with a distribution function. The one that is used to describe the distribution of TF in semiconductor devices, is the Weibull distribution [16]. Another frequently used one is the log-normal distribution, it, however, is less practical to work with, when describing TF for the burn-in of semiconductor devices.

4.2.2 Weibull distribution

The Weibull distribution is often used when describing a failure that is caused by the degradation of the weakest link or when working with reliability of systems, where the whole system can stop to work if one part of the systems fails. This means that when several degradations are present, the TF of the most prominent one, can best be described by the Weibull distribution. In the case of semiconductors the dielectric breakdown of the gate oxide is well described by it, since the whole transistor fails, when a conduction path (at the weakest link) is established in the gate oxide.

The formula for the probability density function (PDF) of the Weibull distribution is:

$$f(t) = \left(\frac{\beta}{\alpha}\right)\left(\frac{t}{\alpha}\right)^{\beta-1} \exp\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right], \tag{28}$$

where t is the time, α is called the scale parameter and β is called the Weibull slope. Depending on β, the shape of the whole density function varies greatly.

The cumulative distribution function (CDF) can be calculated by integrating the probability density:

$$F(t) = \int_0^t f(t')dt' = 1 - \exp\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right] \tag{29}$$

Figure 16 is depicting the PDF and CDF of the Weibull distribution for a scale parameter $\alpha = 1$ and several different slope parameters β.

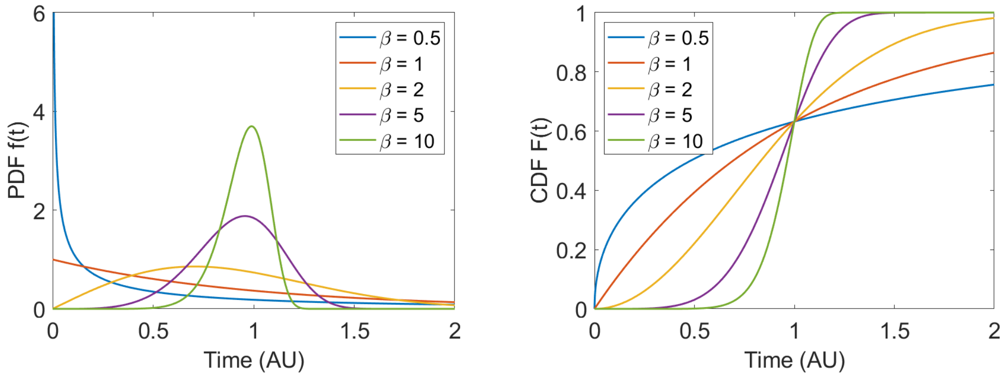


Figure 16: PDF and CDF of the Weibull distribution for several different slope parameters β and a scale parameter of $\alpha = 1$ over time in arbitrary units (AU).

4.2.3 Failure rate and bathtub curve

Regardless of which PDF or CDF is used for describing the amount of failing devices over time, the failure rate λ can be defined along with them.

To put it into words, the failure rate is the amount by which the number of good devices is decreasing or by which the number of failures is increasing. This is in most cases not a linear or constant function of time, but rather depends on when in their lifetime the devices experience degradation to such an extent, that a critical point of degradation is met, which will be at the time-to-failure. It is usually given in units of failures in time (FITs), which are the number of failures per 10^9 device-hours.

Over the course of the mathematical derivation of the failure rate, $M(t)$ will be the number of good devices over time and $M(0)$ is the initial number of good devices at time zero. The rate of change of M with respect to time is given by:

$$\frac{dM}{dt} = -\lambda(t)M(t) \tag{30}$$

where λ is the failure rate. This is the important quantity that wants to be known. Furthermore, one can write the current number of good devices as:

$$M(t) = M(0)(1 - F(t)) \tag{31}$$

with $F(t)$ being the cumulative density of good devices, which can be described by the Weibull distribution, as mentioned earlier.

Combining these the equations yields an expression for the failure rate in terms of PDF and CDF:

$$\lambda(t) = -\frac{1}{M(t)}\frac{dM}{dt} = -\frac{1}{M(0)(1-F(t))}\left(-M(0)\frac{dF}{dt}\right) \tag{32}$$

$$\lambda(t) = \frac{f(t)}{1-F(t)} \tag{33}$$

The equation for λ, when inserting the expressions for the PDF and CDF of the Weibull distributions, yields:

$$\lambda(t) = \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1} \tag{34}$$

Figure 17 shows equation 34 for the same slope and scaling parameters as in figure 16. It is visible that when $\beta = 1$, λ becomes constant, at the value of $1/\alpha$. The higher the scaling parameter, the lower the failure rate. For a $\beta < 1$, λ decreases over time and for a $\beta > 1$, λ increases. These different failure rates can be used to model the expected failure rate over the whole lifetime of a product.

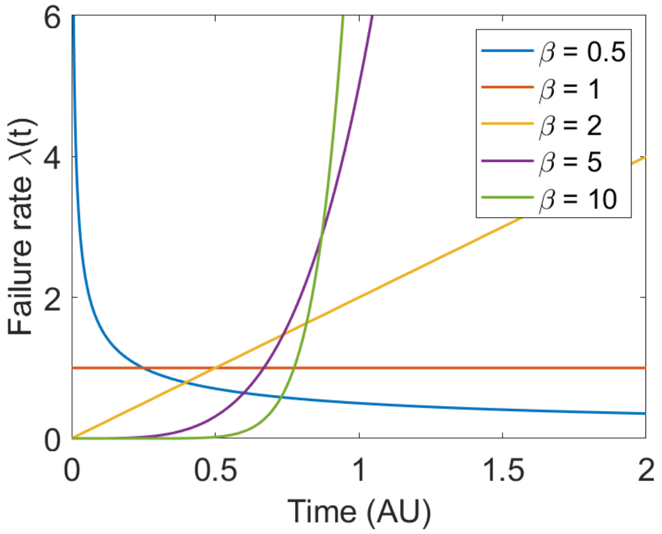


Figure 17: Failure rate λ of the Weibull distribution for several different slope parameters β and a scale parameter of $\alpha = 1$ over time in arbitrary units (AU).

As has been mentioned in this chapter, not all of the devices fail at the same time. This can also be seen in the failure rates above. In order to describe when which device will fail, assuming that all of the devices were manufactured with the same process, one needs to test the devices and observe their time-to-failure behavior. The failure rate of a device (but also of electrical or mechanical products in general) will take the form of the so called bathtub curve.

Figure 18 shows an characteristic bathtub curve, with three distinct regions highlighted in it.

In the beginning there is an initially high failure rate, which decreases over time. This region can be described by a Weibull failure rate with $\beta < 1$ and is called the early life failure rate (ELFR). Towards the end of the life of the devices, the failure rate rapidly increases again, this region is called the wear-out region and can be described by a failure rate with $\beta > 1$. The region between those two is called the intrinsic failure region (IFR) and can be modeled by either a constant failure rate ($\beta = 1$) or as the overlap region between two failure rate models of the ELFR and the wear-out region, in which case $\beta \approx 1$.

The wear-out region will always be there, the only measure which can be taken is to shift it to later times and make the slope less steep. Reliability is therefore only guaranteed until a certain point in time, at which the failure rate is starting to increase again. During the whole IFR region reliability is guaranteed to the costumer, since the number of failing devices is low enough

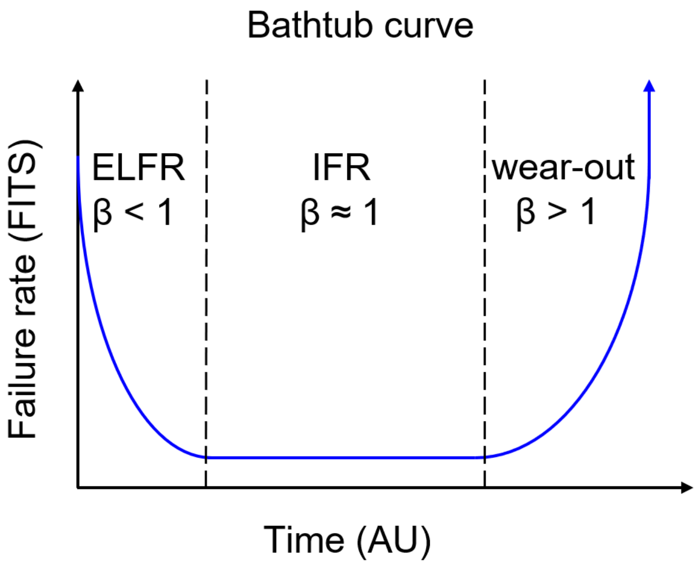


Figure 18: Characteristic shape of a bathtub curve, modeled with three Weibull distributions, resulting in three regions with different slope parameters.

that the few devices which will fail can be reimbursed, however, the ELFR is problematic and needs to be lowered before sending out the devices. One method to do so is the burn-in.

4.2.4 Burn-in

The questions at the end of chapter 4.1 were what the failure rate is and how it can be lowered. The former one of those two has now been answered, while the latter one yet has to be.

In order to lower the ELFR, two approaches can be taken. On one hand the ELFR of the devices can be reduced before sending it to the customer. This can easily happen by operating the devices prior to selling them. The amount of time by which they ought to be in operation, until all the ELFR devices have been sorted out, is given by the device specific bathtub curve, which depends on the manufacturing process. On the other hand, if a process is optimized and every step is well known and works as it should, the ELFR can be decreased without any sorting out of ELFR devices.

Needless to say that the second option is more profitable for the company but also much more difficult to achieve. Therefore, especially in the beginning of the manufacturing process, the time-to-failure is being reduced, by operating the devices. Since the time span, until the IFR is reached, can be quite long (in the range of months to years) a method has to be found to reduce this time. This method is called the burn-in.

During the burn-in the devices get stressed with an increased temperature of 150 °C after already being stressed with an overvoltage. The reason for this is because the time-to-failure depends on the device parameter A_0, which as shown in equation 27 depends on the voltage and temperature.

The laws for accelerated degradation are used to describe how to age devices without operating them for the same amount of time. The used time is then a product of acceleration factors and the stress time.

$$t_{use} = A_T * A_V * t_{stress} \tag{35}$$

Here, A_T is the acceleration factor due to temperature and A_V is the acceleration factor due to voltage [2]. A_T is given by Arrhenius' law [17]:

$$A_T = \exp\left[\frac{-E_a}{k_B}\left(\frac{1}{T_{use}} - \frac{1}{T_{stress}}\right)\right] \tag{36}$$

Where E_a is the activation energy of the failure mechanism, which depends on the material, k_B is the Boltzmann constant, with T_{use} being the normal operation temperature and T_{stress} being the temperature applied during burn-in.

The voltage acceleration is from the Eyring model (often called linear E model). Several voltage acceleration models exist, but for higher temperatures, the Eyring model proved to describe lifetime acceleration the best [18].

$$A_V = \exp\left[\frac{\gamma}{d_{ox}}\left(V_{stress} - V_{use}\right)\right] \tag{37}$$

γ is called the field acceleration factor, d_{ox} the thickness of the gate oxide and V_{use} and V_{stress} are once more the voltage used during normal operation condition and the applied overvoltage, respectively.

Equation 35 to 37 show that the time that the devices need to be in operation can be shortened, by stressing them, in order to reach the IFR.

One goal of the burn-in process is to lower the failure rate to such an extent, that the sold devices comply with the ISO 26262. If the number of devices which are failing due to the burn-in is low, the process can be considered as optimized enough, in order to comply with the ISO 26262 standards even without the burn-in. The burn-in is then not needed anymore. The number of devices which need to go through the burn-in process and the number of devices which are allowed to fail afterwards, can be calculated, for the specific allowed failure rate demanded by the norm. Since these values are statistical predictions on what future failure rates of certain products/devices will be, this prediction has a confidence interval. The confidence interval and the specific number of chips, therefore, vary from company to company and can hence not be mentioned here

Since the activation energy E_a depends on the specific failure mechanism that one wants to activate, as well as the materials that have been used, the following pages will take a closer look on several different failure mechanisms. It starts with the most prominent one, when considering power MOSFETs, which is the breakdown of the gate oxide.

4.3 Defects and failure mechanisms

There is a large number of defects, which can occur in semiconductor devices. Defects can cause failure modes, which can be detected when testing the chip. The underlying failure mechanism describes the effects defects have on the device and how the resulting failure mode comes about. However, not every failure mechanism has a unique failure mode. Several defects can contribute to one failure mechanism, like for example the breakdown of the gate oxide, which will be discussed in greater detail than other failure mechanisms, since it is the most expected one.

4.3.1 Gate oxide breakdown

When discussing the ideal MOS capacitor, two assumptions were made. The first one stated that there were only charge carriers in the semiconductor and in the metal, but not in the insulator. However, it was already mentioned in chapter 3.1.2 that in a real MOS capacitors there are oxide charges in the insulator, which can cause deviations from the ideal characteristics and also enhance an early breakdown of the gate oxide.

Oxide charges: Figure 19 shows a cross section through a MOS capacitor (or the gate of a MOSFET) once more, together with all the different charges, defects and traps that can be present in the insulator.

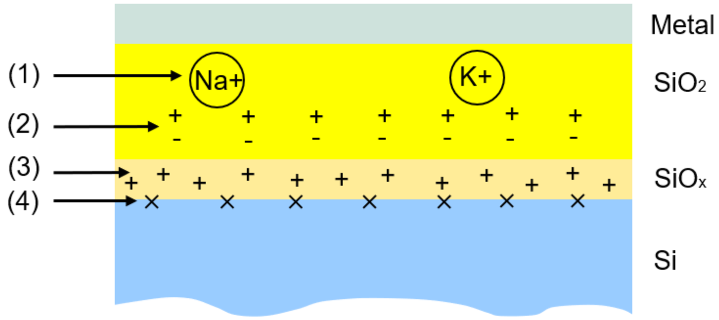


Figure 19: Cross section of gate of a MOSFET, with the respective regions highlighted on the right and (1) the mobile ionic charges, (2) the oxide trapped charges, (3) the fixed oxide charges and (4) interface trapped charges. After [3]

The four different charges that can exist within the insulator are:

- Mobile ionic charges: These are ionized atomic impurities, like Na+ or K+, but also Li+ or H+. They can move around in the insulating material and lead to a change in threshold voltage, since they can act as a permanent bias voltage. Due to an applied field on the outside, they get accelerated.

- Oxide trapped charges: They stem from defects in the SiO_2 which are initially uncharged, but can hold on to charges, either positive or negative. They are distributed throughout the oxide, but can increase in number, when additional charges get trapped or additional defects are generated.

- Fixed oxide charges: When thermally growing a SiO_2 layer on top of a pure Si layer, a monolayer of SiO_x is forming between the stoichiometric, amorphous oxide and the crystalline silicon. The charges are introduced when the incompletely oxidized silicon is formed with low temperature. They are immobile charges that only exist in this small layer.

- Interface trapped charges: As already mentioned, during oxidation a thermally grown, amorphous oxide area is introduced on top of the crystalline silicon. At the interface between the non-stoichiometric SiO_2 and the Si, charge traps can be formed due to defects in the Si crystal or oxidation-induced defects, among other reasons. These traps have states in the bandgap of the silicon semiconductor. Like the fixed oxide charges, they are immobile, but can be charged and discharged [19].

The second assumption which was made during the discussion about ideal MOS capacitors, was that there can be no charge carrier transport through the insulator. This is also not true for real devices, since there are several insulator currents, which contribute to the charge carrier transport.

Insulator current: Insulator currents are of importance to the characteristics of the devices and are also connected to the charge carriers and traps inside and at the interface of the oxide. As there are yet uncharged traps inside the oxide and at the semiconductor-oxide interface, such insulator currents can charge them or even introduce additional charge traps and contribute to the change in device characteristics.

There are two groups of insulator currents. The electrode-limited conduction mechanisms and the bulk-limited conduction mechanisms. While the former one of those depend on the metal-insulator contact and their properties relative to each other, the latter one depends on the electrical properties of the insulator material itself. Statements about the electrical properties like the trap energy level, trap spacing and trap density can be made via the analysis of the bulk-limited conduction mechanisms. The two groups have several different types in them. In the following just a short summary of how the mechanisms work and what they depend on is given. For more detailed information on the subject see [20], [3], [21] and [22]

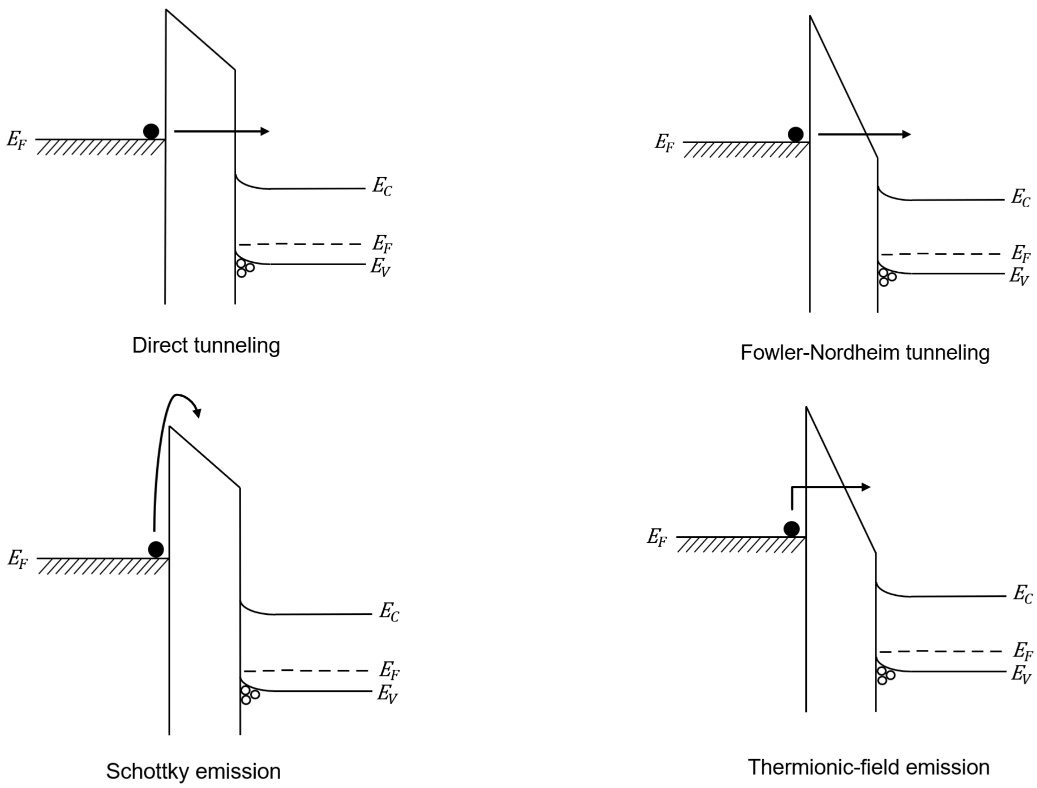
- Electrode-limited conduction mechanisms
    - Direct tunneling: Tunneling is a quantum effect, where the probability of particles, to be in places they classically could not be in, is larger than zero. Therefore, in the case of the gate oxide, an electron in the metal can tunnel through the whole potential barrier of the oxide and reach the semiconductor. Nevertheless, the potential barrier has to be narrow enough in order for the probability to be such high that a significant current can occur. For SiO_2 gate oxides, a thickness of less than 3.5 nm is the border below which direct tunneling becomes the dominant conduction mechanism. For thicker oxides, Fowler-Nordheim tunneling becomes the dominant one.

    - Fowler-Nordheim tunneling: While during direct tunneling, the electron experiences the whole potential barrier of the oxide and tunnels into the conduction band of the semiconductor, during Fowler-Nordheim (F-N) tunneling the electron experiences only part of the potential barrier and tunnels into the conduction band of the insulator. F-N tunneling assumes a triangular potential barrier, which is forming due to the electric field experienced by the potential barrier. Both, direct as well as F-N tunneling is therefore voltage dependent and the larger the applied voltage is, the higher the current flow through or into the insulator.

    - Schottky emission: When electrons receive enough energy, through thermal activation, to overcome the potential barrier from the metal to the insulator, this is called Schottky or thermionic emission. The electrons are then inside the conduction band of the insulator material. This effect is strongly temperature dependent and is affected by charge traps inside the insulator material. Only when the temperature is low enough and Schottky emission is not governing the conduction, tunneling effects can be measured as the dominant mechanism.

    - Thermionic-field emission: When the electrons are thermally activated into higher energies, yet still do not have enough energy to overcome the potential barrier of the insulator like during Schottky emission, they can tunnel through the reduced barrier with a higher probability than during pure F-N tunneling. This is called thermionic-field emission. It is a mixture of thermionic emission and tunneling, therefore, it is temperature dependent, but also voltage dependent.

Figure 20 shows the electrode-limited conduction mechanisms which can occur in real MOS capacitors and MOSFETs.



Direct tunneling

Fowler-Nordheim tunneling

Schottky emission

Thermionic-field emission

Figure 20: Schematic of different types of electrode-limited insulator currents. After [20]

- Bulk-limited conduction mechanisms
  - Frenkel-Poole emission: When electrons are trapped inside charge traps in the insulator material, an applied electric field can lower the coulomb barrier of the trap, making it more probable for the electron to be thermally excited out of the trap and enter the con-

duction band of the insulator. As a result, the Frenkel-Poole emission is most prominent under a combination of high fields and high temperatures. Due to the thermal activation of the electrons, inside the insulator, this effect is often called the internal Schottky effect.

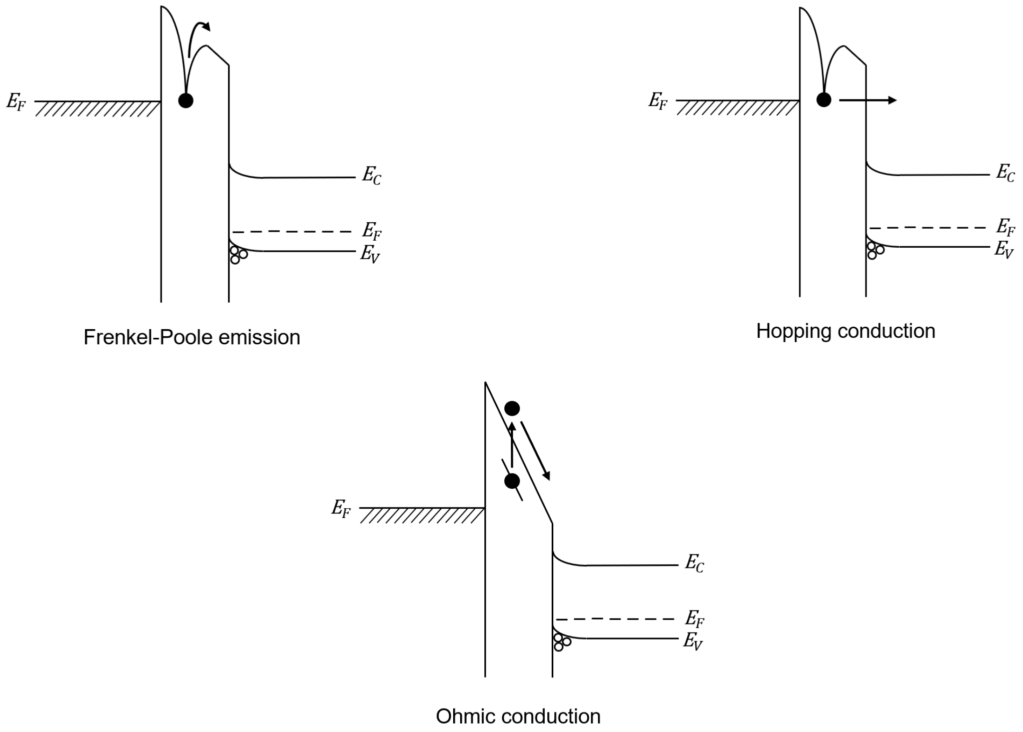
– Hopping conduction: If the electrons inside the traps do not have enough thermal energy to overcome the potential barrier (no Frenkel-Poole emission) they can still tunnel through it. This way the electrons can tunnel from trap to trap, filling the oxide traps and increasing the density of the oxide trapped charges. Hopping conduction has an exponentially decreasing dependency on temperature. This can be explained by the trap energy levels. The trap energy level is the difference between the edge of the conduction band of the insulator and the energy level at the bottom of the trap. This value gets larger with increasing temperature, meaning that at higher temperature deeper traps get activated, which decrease the tunneling probability exponentially.

– Ohmic conduction: Even though the band gap in insulating materials is high, there is still a small number of electrons inside the conduction band, which makes for a small Ohmic conduction through the insulator. This conduction mechanism is linearly dependent on the electric field and can only be observed at very low electric fields. It also increases with temperature.

Figure 21 shows the bulk-limited conduction mechanisms which can occur in real MOS capacitors and MOSFETs.

There are also some other bulk-limited conduction mechanisms, which will not be explained in more detail, because the ones already introduced are enough to explain the gate oxide breakdown. The three missing conduction mechanisms are called space-charge-limited conduction, ionic conduction (coming from the mobile ionic charges) and grain-boundary-limited conduction. More information on all of those can be found in [20].

If those insulator currents interact with the charge traps inside the insulator, they can charge and discharge the traps and current can flow more easily. If enough of them are charged, a conducting pathway can be established between the metal and the semiconductor, through the insulator, which results in a breakdown of the gate oxide. This is the theory behind the failure mechanism of the oxide breakdown. The corresponding failure mode, which can be detected, is an increase in current at the gate contact in the 'on' and 'off' state, as well as an increase in drain current in 'off' state only. This current



Frenkel-Poole emission

Hopping conduction

Ohmic conduction

Figure 21: Schematic of different types of bulk-limited insulator currents. After [20]

is considered as leakage current when talking about it at an engineering level. Chapter 5 explains how and at which pins the leakage current can be measured on the BIRD and what its significance is.

Devices with too many charge traps and defects in the insulator material experience a breakdown of the gate oxide earlier and therefore contribute to the ELFR. Naturally, good devices, can experience this too, when coming into wear-out, however, contributing to the increase in charge traps is an effect called hot carrier injection (HCI). This effect can not only contribute to an earlier oxide breakdown, but also to degradations which affect parameters like the threshold voltage or the on resistance.

4.3.2 Hot-carrier-injection

HCI is a mechanism more observed in MOSFETs, rather than in MOS capacitors, because it requires large electric fields that can accelerate electrons to high energies. In order to understand the HCI, a short discussion about some properties of real MOSFETs is necessary.

Properties of real MOSFETs: When plotting the increase of the drain current, as a function of the gate to source voltage, the resulting curve, for a fixed drain voltage is a linear function of the gate voltage, which starts to rise at $V_{GS} = V_T$. This relation can be easily derived when looking at equation 14.

Real devices, however, have a non-zero drain current already before this condition is met. The reason therefore is, that when the gate bias is just below the threshold voltage, the device is in weak inversion. That means that there is not yet a full inversion channel formed, but there is still some current to the drain region. This current contributes to the drain leakage current when the device is still in 'off' state.

Another effect which results in deviations from the gradual channel approximation is happening at higher gate voltages. Even if the gate voltage is increasing, the drain current will eventually settle at a certain value. This is due to the fact that the carrier mobility is not independent of the electric field. The mobility becomes lower at high fields until it settles. An explanation for this is that the electric field gets significantly higher near the drain region of the MOSFET. This high field accelerated the electrons to the interface between the silicon and the gate oxide, where they can excite optical phonons. This only happens at high enough energies which results in more scattering of the electrons and a saturation of the drain current.

Figure 22 shows the mentioned effects in real MOSFET devices in a qualitative way. The blue curve shows the ideal drain current increase, when following the linear term of the gradual channel approximation. The red, dotted lines indicate the deviations from the gradual channel approximation due to the subthreshold current and the mobility decrease. Further readings on these two effects can be found in [3] and [4].

Chapter 3.2.1 states 4 assumptions about MOSFETs in order to derive the current-voltage characteristic of the devices. One of them (assumption 3) was that the electron velocity v_n is constant and independent of the electric field, which is true for low electric fields. As has already been explained, at higher fields the mobility decreases with the rising electric field. One other phenomenon that can happen, is that electrons get accelerated into high energies

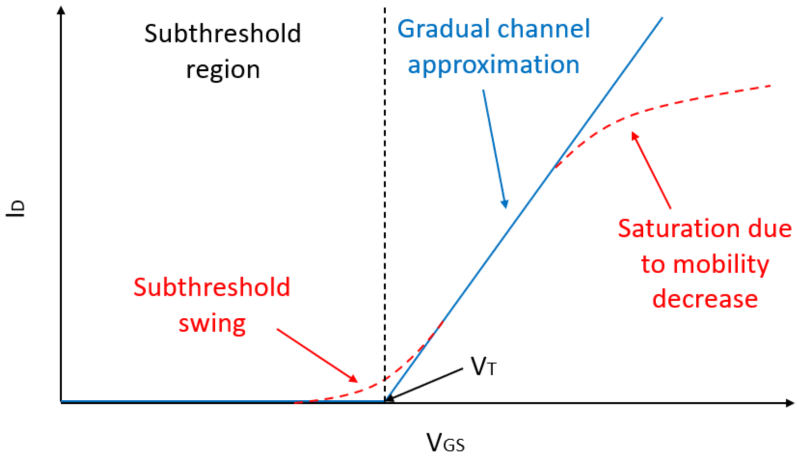

Subthreshold region
Gradual channel approximation
ID
Saturation due to mobility decrease
Subthreshold swing
VT
VGS

Figure 22: Drain current as a function of the voltage between source and gate. The blue line is the drain current when using the gradual channel approximation in the linear regime, while the red dotted lines are representing the curves behaviour in real devices.

due to a locally high electric field near the drain region of the MOSFET.

As has already been hinted, real devices do not necessarily have low electric fields and therefore also do not have a constant v_n or electron energy. Especially near the drain region, the electric field of the device increases due to pinch off. At pinch off, the gradual channel approximation that was used to derive the drain current, does not apply anymore, due to the violation of the assumptions.

The part of the channel that is pinched off is fully depleted and the electric field would diverge, when further using the gradual channel approximation. In reality it does not diverge, but it does, however, increase in orders of magnitude. In order to calculate the electric field at pinch off and in consequence the correct voltage across the channel and the correct drain current, the channel length modulation is a model which can be applied. For further reading see [4].

During HCI, electrons are accelerated to high energies by the increasing electric field near the drain region. Two mechanisms can be distinguished. On one hand the field-driven HCI and on the other hand HCI at low voltages, where multi-electron effects play a role. The latter one of those is only relevant in devices with small channel length that are being operated at voltages below 1 V, which does not apply for the devices on the BIRD.

Regarding field-driven HCI (which is a single-electron effect), the accelerated electron can damage the passivized Si bonds at the Si-SiO$_2$ interface. These passivized bonds are the SiH bonds which are introduced before the gate oxide is grown, in order to minimize interface traps due to the lattice mismatch between the Si and the SiO$_2$.

However, when the hot electron is accelerated to energies greater than the bonding energy of these bonds (1.5 eV [23]), the H atoms get detached and a dangling bond is left, which acts as an interface traps. As mentioned in 4.3.1, these interface traps can capture electrons from the insulator currents or the hot electron can get caught itself. [24]

When additional traps are introduced due to HCI more charges can stick at the interface, resulting in a 'permanent' bias voltage. This permanent bias is the reason why the threshold voltage V_{th} shifts.

Figure 23 is showing the principles of the HCI mechanism at the drain side of a basic lateral MOSFET.

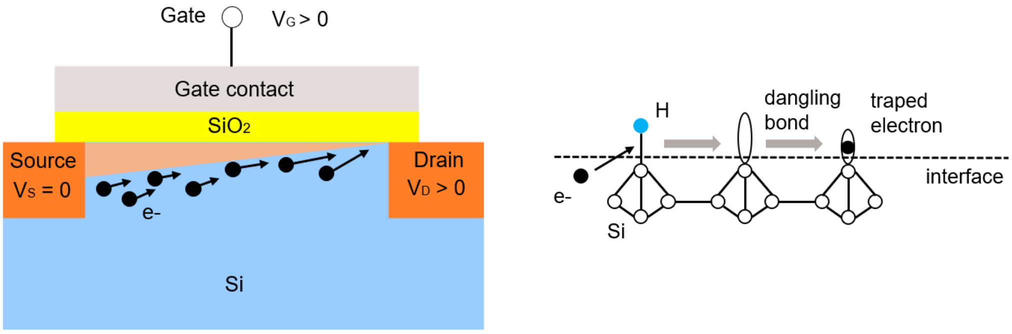


Figure 23: Sketch of the working principles of the HCI mechanism. On the left, the electrons gain energy due to the increasing electric field in the pinched off region of the channel. On the right the process of hot electrons breaking up the SiH bond and the resulting dangling bonds getting charged is shown. After [25]

The hot electrons can also jump the potential barrier of the Si-SiO$_2$ interface (3.1 eV [3]) and contribute either to the gate current (only small compared to other insulator currents) or introduce additional oxide traps. As explained in 4.3.1 these traps, when charged, can form a breakdown path and lead to the breakdown of the gate oxide.

Another region where the HCI can effect the device performance is at the STI, which is used as a drain extension. At the edge of the STI the hot electrons can also introduce interface traps at the interface between the STI and the Si in the drift region. The traps become charged and as more and more of the traps become charged, an electric field which accelerates the electrons additionally builds up. This increases impact ionization at the bottom corner of the STI, which results in an increase in an off state leakage current. As shown in [26] the leakage current can increase up to three orders of magnitude after an operation time of only 15 hours, making HCI a contributor to the ELFR once more.

In order to prevent this degradation from happening to such an extent, different LDMOS with STI drain extension geometries were investigated.

4.3.3 Short of the shallow trench isolation

A similar effect to the breakdown of the gate oxide is a short of the STI. As has already been introduced in chapter 3.4.2, the STI in power MOSFETs like the LDMOS, which are on the BIRD, is used as a drain extension in order to increase the drift region of the device. The other application for the STI is the isolation between the transistors itself or between the body and the drain/source contact.

The steps involved in the manufacturing process of the STI include a photolithography, followed by a chemical etching process in order to make the trench. Photolithography in the semiconductor business is a process where light is emitted onto certain parts of a photoresist. This photoresist changes its chemical properties when hit by light, which makes it possible to dissolve it in a chemical solution. The parts of the photoresist, which were not exposed to the light still have the photoresist on them after the developing. It acts as a shield for the areas underneath, which should not be etched away, since it is resistant to the etching process. If some of the photoresist has not been eliminated during the lithography, the chemical etch can not fully etch the trench in that region. This is called micromasking and can result in a local thinning of the STI later on [27].

That way, especially when the STI has to withstand large voltages, a breakdown of the STI can occur and electrically connect regions in the device that should be isolated from each other, for example the gate and the drift regions of the LDMOS. The failure mechanism in this case would be similar to the breakdown of the gate oxide, since the STI material is also an insulating dielectric, resulting in the same failure mode. Reasons for the defect can, for example, be particles on top of the photomask or directly on the wafer, which block the light and prevent the photoresist from being developed.

Figure 24 shows the cross section of the same LDMOS which was shown in figure 14, however, a micromasking defect in the STI was added, which thins out the insulator material locally.

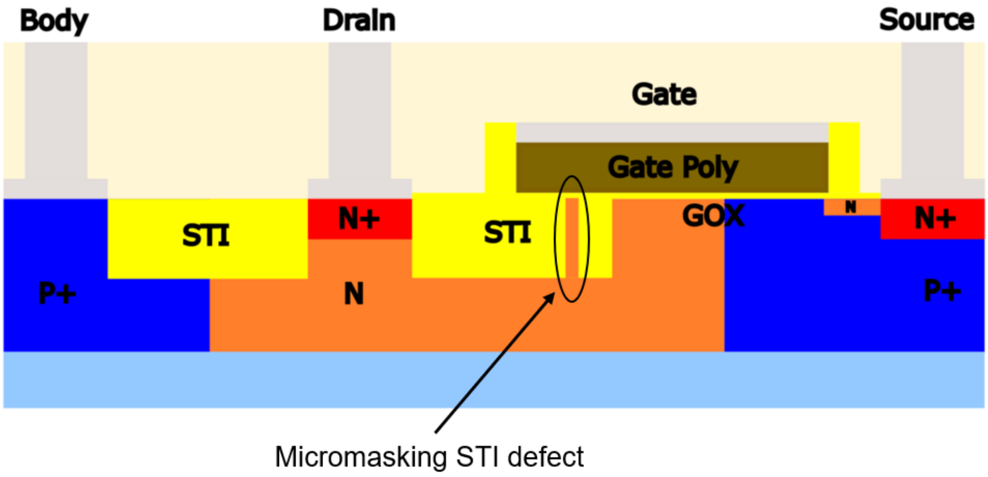


Figure 24: Cross section of an NLDMOS with a defect in the STI due to micromasking.

5 Electrical characterization of the BIRD

In the previous chapters, an overview about the BIRD as well as its semiconducting devices was given, along with a dedicated chapter about the theory and physics of transistors. Chapter four then contained the standards that need to be met and three failure mechanisms were introduced. Everything, that needs to be known in order to understand the electrical characterization of the BIRD has been established. The following chapter focuses on the parameters that were tested. The discussion will include the reasoning, as well as the order in which they were tested. This follows a description about methods on how to evaluate large numbers of electrical data. Understanding these methods, together with the theory about transistors and applying them to the electrical parameters that are tested, gives rise to the last part of this chapter, namely, the production limit setting. The production limits are used for screening out malfunctioning chips and monitoring the performance of the production line.

5.1 Electrical parameters and testflow

The way to test the BIRD automatically is by means of a testprogram (TP). Depending on the tester system and company standards, different programming languages can be used to code the TP. This TP is written in Visual Basic For Test (VBT), a coding environment embedded in the IG-XL data tool from the company Teradyne. IG-XL itself is within the Microsoft Excel software.

The TP gives instructions to the tester, a machine, which by contacting the chip applies currents or voltages to the chip in order to read out and measure other electrical parameters like current, voltage or resistance.

There are many ways a TP can look like and it always depends on which tester or programming language is used and what parameters one wants to test in the chip. The big advantage of the BIRD TP is that, very much like the design of the chip itself, it is written to be reusable for future burn-in studies as well. This can however only be realized, when the general layout of the circuits of the future BIRD chip stays more or less the same. In this case the actual transistors which are on the BIRD can be different ones and the TP would still be usable with only minor adjustments.

5.1.1 Electrical parameters

There are a variety of tests that are integrated in the test program. Most of them are used to ensure proper testing and reading out of the electrical parameters necessary for characterizing and evaluating the functionality of the devices on the chip. To get an overview, table 3 provides a list of the parameters, together with their unit and the electrical quantity that is used to generate the wanted data. It also contains a column with the category of the test. This gives information about whether the test is used for ensuring proper measurement (contact tests) or for measuring the performance of the devices (parametric tests).

To understand this in more depth, each test is being explained in more detail in the following paragraphs:

Kelvin: The kelvin test is also called four-point probes method or four-terminal sensing. To perform the kelvin test, two force and two sense lines need to make contact with the pins of the chip. The force lines are connected to a DC current and the sense lines are connected to a voltmeter. By applying a constant current via the force line, and measuring the voltage drop on the contacted

Testname	**Unit**	**Quantities applied by tester**	**Category**
Kelvin	Ω	Current	Contact
Continuity	V	Voltage/Current	Contact
Chip-ID	*digital*	Voltage	-
Leakage current	nA	Voltage	Parametric
Stress	nA	Voltage	-
Capacitance	pF	Voltage	Parametric
On resistance	Ω	Voltage/Current	Parametric
Fall time/rise time	μs	Voltage	Parametric
Delta continuity	mV	Voltage/Current	Contact

Table 3: Electrical parameter whose values are measured, together with their respective units and physical quantity the tester uses to measure them. The category column shows which tests are contact test, parametric test or neither.

pin via the sense line, the resistance of the pin can be calculated.

If the resistance is small, a good contact has been established and testing can proceed. A high resistance could indicate that the pin is either not properly connected or a large voltage drop occurs at the pin, which may harm it. Therefore a maximum voltage drop is set in the tester above which a relay opens that prevents additional current to flow through the pin. This is called clamping. Further reading on the Kelvin test can be found in [28]

Continuity: This test is performed by forward biasing one of the ESD diodes with a fixed current. The voltage drop on the diode is measured. If the voltage drop lies within a certain range around the expected value, a good contact between the device and the tester has been established. Therefore, the continuity test is also a contact test. For example a bad interfacing between the device and the test equipment. However, the problem can also lie within the fabrication process of the chip. For instance a loosely attached or broken bond wire results in an open circuit, which cannot be detected by the Kelvin test.

Chip-ID: These tests store the chip-ID by writing it into the storage in the digital part of the chip. This happens by applying different voltage signals to the pins dedicated for the digital part. One part of these tests is to write the chip-ID into the storage, another one is the read it out again. Once written, the ID can not be changed, since the digital part includes a one time programmable (OTP) digital storage. It can, however, always be read. The used OTP is an eFuse OTP, for further reading on how such an OTP memory works, see [29].

Leakage current: This series of tests measure the leakage current at the gate, drain and source (if available) pins. A high leakage current, indicates that there is one or several segments in the current path, which are lower in resistance than expected. The breakdown of the gate oxide, as discussed in chapter 4.3.1 is one of the events which can increase the leakage current.

Several leakage tests are conducted on each device. On the NMOS for example, the leakage current is measured when the transistor is on and off. While enable is on low, which means that the transistor is switched off, the gate and drain leakages are measured. When switched on, only the gate leakage is measured, the drain leakage not, since the device is in saturation mode and conducting anyway. Since the gate leakage is measured in both, on and off states, the difference between those two can be calculated. This delta value is ideally zero, since that would mean that during usage, the leakage current is not increasing.

In order to measure the drain leakage values for instance, a fixed positive voltage is applied to the gate pin of the NMOS, while the enable line is at low, so that the transistor is still in the cut-off region. If one applies a positive voltage to the drain pin, a current can be measured on the drain pin, which should be as small as possible.

Stress tests: During the stress tests, half of the devices, which have the stressed device area are being exposed to an overvoltage, specific for each device. During the stress tests the leakage current is also measured. These tests were merely implemented to have a test block inside the TP that indicates the stressing, the data are not being evaluated. Therefore they are neither parametric nor contact tests.

Leakage current post: After the devices were being stressed, the leakage current is measured again, to see if the overvoltage stress applied to the devices affected the leakage, e.g. by damaging the gate oxide of the transistor. The procedure is the same as for the initial leakage current tests before the stress was applied.

Capacitance: This test measures the capacitance of the MOS capacitor. To do so an AC voltage signal is sent into the device and by measuring the phase shift of the applied signal, compared to the measured one, the capacitance of the device is measured. However, since there are several parasitic effects, due to the other electrical components in front of each device for example, at first a calibration measurement is done, while the capacitor is switched off. That way only the parasitic effects are measured. The calculated parasitic capacitance is then subtracted from the final measurement.

On resistance: The on resistance tests measure the resistance between drain and source of the transistor when it is in the saturation regime. Therefore at first the gate voltage is set to a dedicated value and on the enable line the transistor is switched on. A fixed current will be forced from the drain pin and the drain voltage will be measured. The on resistance is the voltage between drain and source divided by the drain current. For a specified gate voltage, the resistance is known due to simulations, therefore the values are expected to be close to those simulation values.

Fall/rise time: The fall and rise time are also two parametric values that are being measured. They are defined as the time the transistor needs, to establish a desired voltage between drain and source after switching the transistor on/off. To be specific, the rise time is the time the transistor needs to go from 10 % to 90 % of the desired V_{DS}. The fall time on the other hand is the time it needs to go from 90 % back to 10 % again.

$$t_{rise} = t(V_{DS,90}) - t(V_{DS,10}) \tag{38}$$

$$t_{fall} = t(V_{DS,10}) - t(V_{DS,90}) \tag{39}$$

A shorter time is equal to a steeper slope of the voltage curve and also to a faster switching on/off process. The switching performance of a MOSFET is an important characteristic, because it also gives information about how much power is dissipated during switching. The longer the fall/rise time, the more power is dissipated.

Delta Continuity: At the end of the TP, the continuity tests are performed again in the exact same way as they were at the beginning of the TP. The difference between the first and the second continuity test is calculated and data logged. In the ideal case, the difference is very closely to zero, which means that the quality of the contact did not change throughout the measurements. The reason, there might be a shift of voltage in the second continuity test, relative to the first one is that during the measurements mechanical vibrations could loosen the contact.

In general all these tests need to be done at least once on all the devices on the chip, which results in a TP that has 400-500 tests in total. This was a short explanation about the most important tests included in the TP. There are also tests that were not mentioned here, because they are less important when specifying the devices. They serve the purpose of ensuring and monitoring the proper execution of the TP.

5.1.2 Testflow

Now that the tests have been introduced and explained, it is necessary to mention that these tests are run more than once. Before assembling the chip, which means before cutting the wafer, gluing the chips inside of their packages and bonding them, the TP tests the whole wafer one time (that means over every chip) at an environment temperature of 150 °C. This process is called front end wafer testing, with the front end (FE) being everything that happens before assembly. After assembly the chips get to back end (BE) testing, where they are again tested. This time, however, at -43 °C which is called cold insertion and afterwards again at 150 °C which is called hot insertion. Testing at different temperatures has advantages for screening and activating failures, since some effects can only be detected at a certain temperature range. [30] All the insertions, together with their respective tests are called the testflow. Also the order in which the tests are performed on each device are included in the testflow. An overview of this can be seen in figure 25. The different colors of the tests indicate different purposes. The green tests are contacting tests, whereas the orange tests are parametric tests. Blue is the capacitance test of the MOS capacitor, which is also a parametric test and the pink ones are the tests necessary for storing and reading the chip-ID.

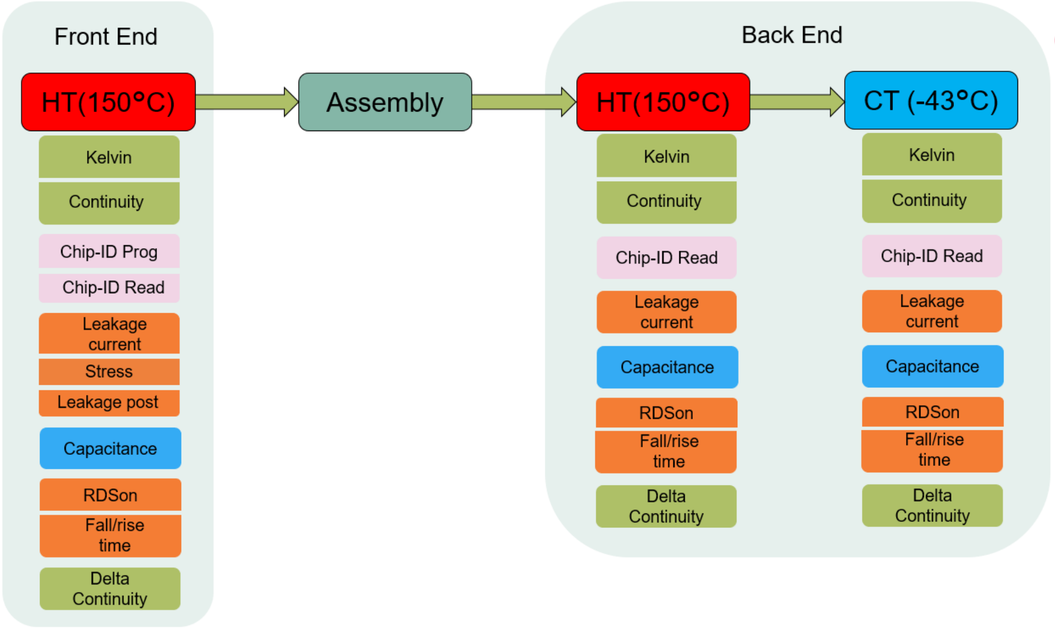


Figure 25: Testflow of each BIRD, before the burn-in starts.

One can also see that there are slight differences in the tests that are performed between FE and BE. This has to do with the different device sizes which were mentioned in chapter 2. The electrical stress that is applied on half of the devices is only applied during FE wafer testing. Therefore, the TP is slightly longer, since the leakage current is measured once before and once after stressing the devices. Also the Chip-ID Prog tests are not included in BE anymore, because the chip-ID is only written in FE, when the chips are still on the wafer. In BE, the tests performed during the hot and cold insertion are the same ones.

5.2 Basics for electrical data evaluation

Besides giving an overview of the tests that are being done on the BIRD and what they measure, it is important to mention that during the course of a BI study more than 100.000 chips are being tested several times. It is expected of the arising data to behave in a way that statistical processes can describe it. One important distribution which can be applied to the data is the normal distribution.

Normal distribution and cumulative distribution function

The measured data is often times normally distributed, following the Gaussian probability density:

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{40}$$

Furthermore, the cumulative distribution function of the Gaussian probability density is given by the integral:

$$F(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt \tag{41}$$

This integral can be solved by substitution, which leads to the expression:

$$F(x) = \frac{1}{2}\left(1 + erf\left(\frac{x}{\sqrt{2}}\right)\right) \tag{42}$$

with $erf(x)$ being the Gaussian errorfuction:

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-\tau^2} d\tau$$

In 26 on the left side, the gaussian cumulative distribution function is depicted. If the y-Axis is rescaled to represent quantile probabilities, the sigmoid shaped curve of the cumulative distribution function becomes a straight line, which has multiple advantages during the evaluation of the data. If there are a lot of tests, which should ideally all be normally distributed, it is much easier to see deviations from desired distribution if it is represented as a straight line.

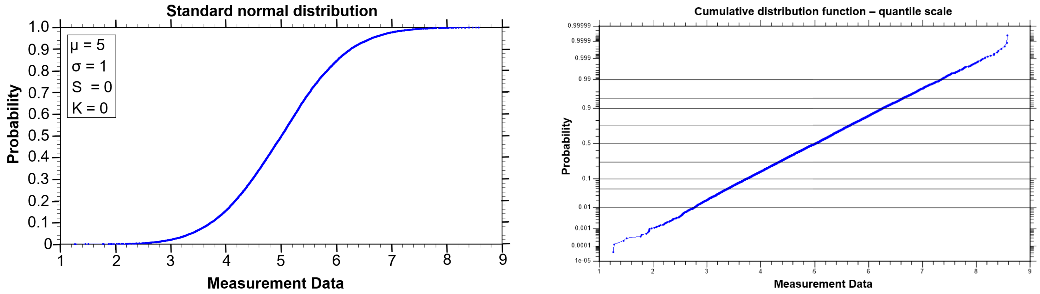


Figure 26: Gaussian cumulative distribution function with a linear scale on the left and with a quantile scale y-axis on the right. Example data is a simulated normal distribution with $\mu = 5$ and $\sigma = 1$.

While the plot on the left has a linear y-axis, the y-axis of the right plot is scaled logarithmically, centered around 0.5. The slight deviations from the perfect normal distribution towards the upper and lower end of the distribution is better visible in the right plot, which is why this representation is chosen. The data used, was generated with the MATLAB normrnd function and is not from a real measurement.

The measurement data, gathered by the TP is only normally distributed, when every component that contributes to the measurement and to the manufacturing process is also normally distributed. If there is, for example, a systematic effect that shifts the upper end of the distribution to higher values, it is caused by a distinct mechanism. Often times this mechanism is not investigated, because the data still is close enough to an ideal normal distribution or is not in danger of violating any specifications. Sometimes, however, when the mechanism is of interest, the shape of the distributions can help to determine the origin of the unwanted mechanism.

Two quantities of general data distributions, which determine the shape of them, are for example the skewness and the kurtosis. The skewness measured how assymetrical the distribution is and the kurtosis measures if the distribution is heavy or lightly tailed.

Those two quantities also have mathematical definitions. For the skewness it is the adjusted Fisher-Pearson coefficient:

$$S = \frac{\sqrt{N(N-1)}}{N-2} \frac{\sum_{i=1}^{N} (Y_i - \mu)^3}{N\sigma^3} \tag{43}$$

N is here the number of data points and Y_i the data itself. The first fraction is the sample size adjustment. It approaches 1, as N gets large.

If the data is normally distributed, the skewness is zero. When the data is skewed to the left, S is negative, if the data is skewed to the right, S is positive.[31]

The kurtosis of a data distribution is given by the expression:

$$K = \frac{\sum_{i=1}^{N}(Y_i - \mu)^4}{N\sigma^4} \tag{44}$$

The normal distribution has a K of 3. Because of this, the following definition is often used:

$$K = \frac{\sum_{i=1}^{N}(Y_i - \mu)^4}{N\sigma^4} - 3 \tag{45}$$

In this case, a positive K means a heavily tailed distribution and a negative K means a lightly tailed distribution. [32]

One other thing that can appear in real measurement data, is a double distribution. This is given when one subset of the data is normally distributed around a different mean value than the other subset. Of course there can also be more than two subsets, in that case it would be multi distributed data.

In figure 27 four histograms of 12000 MATLAB generated data samples are plotted. On the top left of each plot, the mean values, standard deviations, the skewness factor S and the kurtosis factor K are found. The MATLAB function pearsrnd was used to generate the sample data.

One can see what effects the different values for skewness and kurotsis have on the data distributions. The data on the top left is standard normally distributed, on the top right the data is skewed to the left. The bottom left data has a high kurtosis and the bottom right data consists of two normally distributed datasets with different mean values.

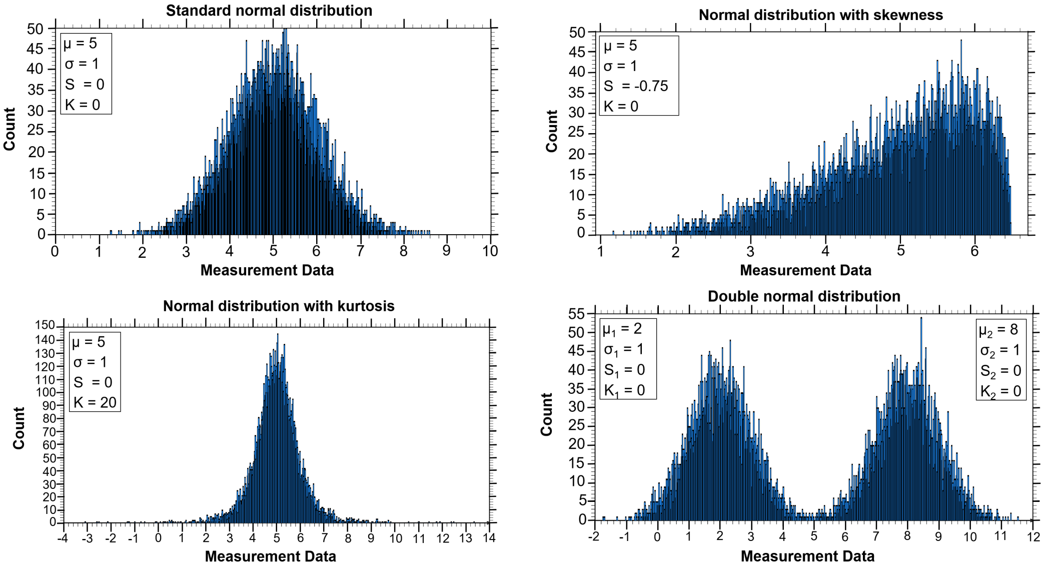


Figure 27: Histograms of 12000 data samples with different values for μ, σ, S and K.

Upon rescaling the y-axis, the quantile representation of the cumulative distribution function is plotted in figure 29, again with the same data as in the previous plots. The effects of skewness, kurtosis and double distributions in this data representation are clearly visible.

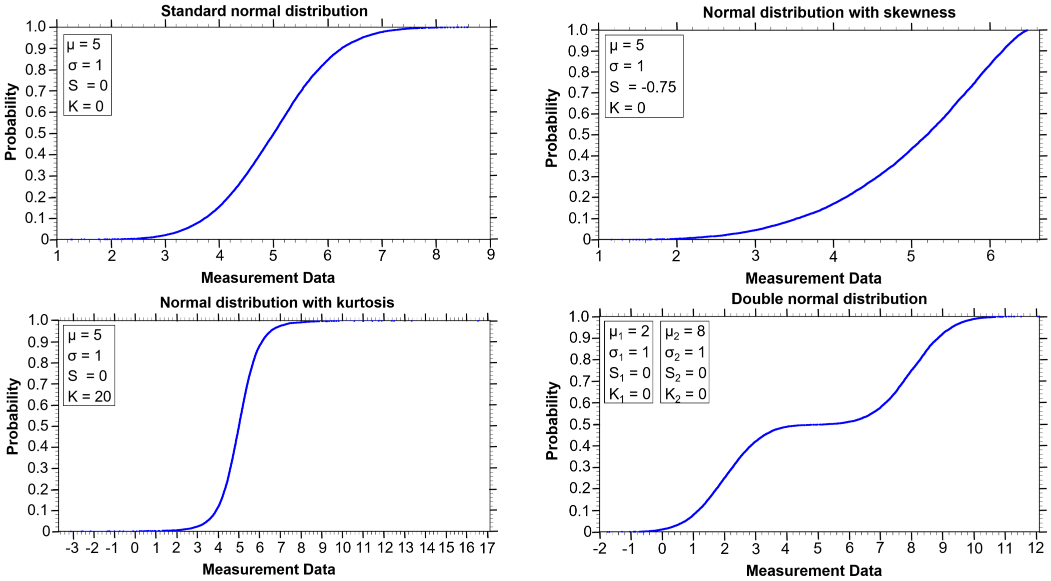


Figure 28: Cumulative distribution functions with linear y-scales of 12000 data samples with different values for μ, σ, S and K.

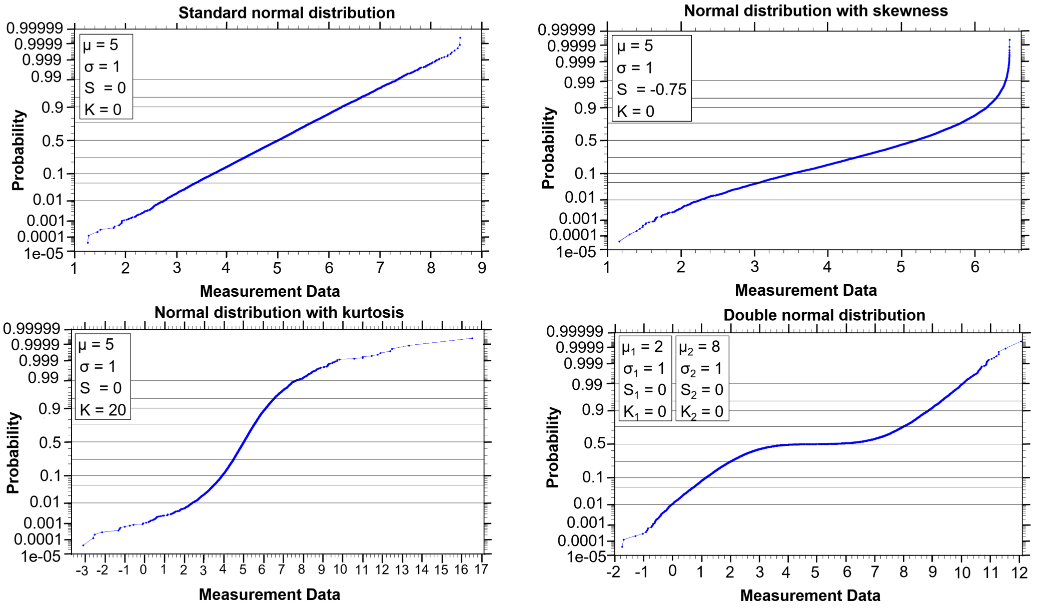


Figure 29: Cumulative distribution functions with quantile y-scales of 12000 data samples with different values for μ, σ, S and K.

$6\,\sigma$ approach:

Now that it is established that the data should be normally distributed, what deviations exist and that we use the quantile representation of the cumulative distribution function, it needs to be clarified on how to assess, if measured data is "good" or "bad". One way to do this, is to introduce limits between which the measured data should lie. But how are the limits chosen, since the data is normally distributed, and there is a finite probability that a value can lie far off the mean value?

For BIRD the so called $6\,\sigma$-approach was used. In general, six sigma is a whole management system which aims for process optimization.[?] In the case of BIRD, the quality of the manufacturing process is improved, by identifying the causes of defects and minimizing variability. Here, a six sigma process is one in which 99.99966% of all opportunities to produce some feature of a part, are statistically expected to be free of defects. Therefore, one can introduce limits that fulfill the $6\,\sigma$ criterion. In other words, $6\,\sigma$ is used as a statistical tool during the BIRD BI study.

The underlying statistics of a $6\,\sigma$ process is that it corresponds to a specific defects per million opportunities (DPMO) value. The DPMO is defined as:

$$\mathrm{DPMO} = \frac{\mathrm{TND's}}{\mathrm{N} \cdot \mathrm{NDO's}} \cdot 1000000,$$

where TND's stands for the total number of defects found in a sample, N is the sample size and NDO's is the number of defect opportunities per unit in the sample. This is directly linked to the a percentage of devices without defects. For a $6\,\sigma$ process the DPMO value is 3.4. In a project like the BIRD, which is not a product with costumer specifications, the DPMO is the limiting factor of which limits to apply.

However, one needs to consider, that for every measured, normally distributed quantity, a long term shift of $1.5\,\sigma$ is assumed, which means that the mean value of the distribution shifts with $\pm 1.5\,\sigma$. The amount of the shift is an industry standard, which was observed in several processes over the years, by many companies. Process deterioration and many subtle performance changes in the production, which are too hard to control and monitor, are the cause for the $\pm 1.5\,\sigma$ drift. This shift naturally results in a higher amount of devices overstepping the 6σ limit. Therefore it is necessary, to take it into account upon initially setting the limits.

C_{pK} and C_p process capability indexes:

The process capability indexes (C_{pK} and C_p) are a statistical method of measuring the ability to produce an output within specification limits. The set of limits used here are ensuring $6\,\sigma$, even including the $1.5\,\sigma$ drift. The formula for the C_p is:

$$C_p = \frac{USL - LSL}{6\sigma} \tag{46}$$

and the formula for the C_{pK} is:

$$C_{pK} = min\left[\frac{USL - \mu}{3\sigma}, \frac{\mu - LSL}{3\sigma}\right] \tag{47}$$

USL stands for upper specification limit and LSL stands for lower specification limit. As can be seen from the mathematical expressions, the C_p takes the whole range between the specification limit and divides it by the $6\,\sigma$. On the other hand the C_{pK} takes either the range between USL and the mean value or between LSL and the mean value and divides it by only $3\,\sigma$. The smaller one between these two is the C_{pK}. The difference between these two indexes is when which should be applied.Now depending on which type of data and limits is at hand, either the one or the other is better applied. If the data is not expected to be centered between the specification limits, the C_{pK} is of greater use, because the C_p overestimates the process capability. It is also sensitive to mean value drifts. These expressions for the C_{pK} and C_p are only valid if the data is normally distributed, for other distributions the expressions can be found in the DIN ISO 22514-2 norm.

With the introduction of the C_{pK} and C_p, there is now a direct link between the limits and the $6\,\sigma$ approach. In order to find the appropriate limits for the tests in the TP, one can look in table 4. One has to search for the row with the wanted DPMO and sigma level. The wanted DPMO for the BIRD is 3.4, which corresponds to a $6\,\sigma$ level, which further corresponds to a long term C_{pK} of 1.5. Together with the equation for the C_{pK}, the limits can now be calculated for each test.

This method, however is only appropriate, if there are no specification limits at hand. Otherwise the limits are fixed and the process has to be optimized until a C_{pK} of 1.5 is reached.

When the limits for the individual tests are implemented, every chip, which

Sigma level	**DPMO**	**with defect**	**without defect**	**Short term C_{pk}**	**Long term C_{pk}**
1	691462	68%	32%	0.33	-0.17
2	308538	31%	69%	067	0.17
3	66807	6.7%	93.3%	1.00	0.5
4	6210	0.62%	99.38%	1.33	0.83
5	233	0.023%	99.977%	1.67	1.17
6	**3.4**	**0.00034%**	**99.99966%**	**2.00**	**1.5**
7	0.019	0.0000019%	99.9999981%	2.33	1.83

Table 4: Relation between sigma, DPMO, parts with and without defects, as well as their respective long and short term C_{pk} values.

violates these limits is taken out of the process line. This is called screening.

5.3 BIRD electrical data evaluation

After establishing what is tested and how to assess the measured data, the last step is to combine those two. At first a certain amount of BIRD chips are tested with the use of the TP. They experience the testflow as presented in figure 25. The measured data from each test is plotted using the quantile representation of the cumulative distribution function, in order to assess if the data is normally distributed. If that is the case, limits can be introduced which satisfy the $6\,\sigma$ criteria. Non normally distributed test data, needs to be assessed as to whether it is a problem or not. The tests which are the most interesting to characterize the new transistors on the BIRD, are the parametric tests, namely leakage current, on resistance and rise/fall time.

Before the actual BI study begins, several BIRD chips are tested with the BE TP, in order to check a first set of data in advance. This way, potential TP bugs can be detected and corrected. Another benefit is that an initial frame of reference regarding the mean value and the standard deviation of the data taken by each test can be established. 99 BIRD chips were tested twice, first in a hot insertion (150 °C), then in a cold insertion (-43 °C), during this initial step. The devices on these chips did not undergo FE testing, so no malfunctioning devices have been screened yet.

Since there are several hundred tests in the TP and plotting each distribution would defeat the purpose of this chapter, just a small collection of representative data has been chosen for the following plots.

Rise time NDMOS-60V: In figure 30 on the left hand side, the two rise times of the NDMOS-60V device are depicted. The data for both BE insertions have been plotted in the quantile representation of the cumulative distribution function. The values have been normalized to the highest occurring rise time of the hot insertion.

This is an example of nicely normally distributed data of a parametric test, during both insertions. The limits (USL & LSL) are calculated by using a long term C_{pK} of 1.5, which corresponds to a DPMO of 3.4 and a sigma level of 6, as discussed in chapter 5.2. One can see, that there are no limit violations in both data sets.

The rest of the rise and fall time measurements yielded similar results, with no limit violations and normally distributed data.

Kelvin test NDMOS-40V: On the right hand side of figure 30 the measurement results of both insertion of the kelvin measurements of the NDMOS-40V have been depicted. The values were again normalized to the maximum value of the hot insertion. This is an example of non ideally distributed data, since, especially the hot insertion is tailing towards higher values.

The shift between hot and cold due to temperature is an anticipated effect, since the resistance of a conductor rises with temperature. The kelvin tests, however, do not have separate limits for the hot and the cold insertion. As a matter of fact, all kelvin tests have a LSL of $0.1\,\Omega$ and a USL of $1.52\,\Omega$, which is just below the clamping voltage mentioned in the test description in chapter 5.1.

The reason for these limits, which do not have a C_{pK} of 1.5, is that outliers in the kelvin measurements are not necessarily a reason for screening the chip. Because the kelvin resistance is a contact measurement, as long as the contact is made, the measurement can be done normally, even if certain pins might be higher in resistance.

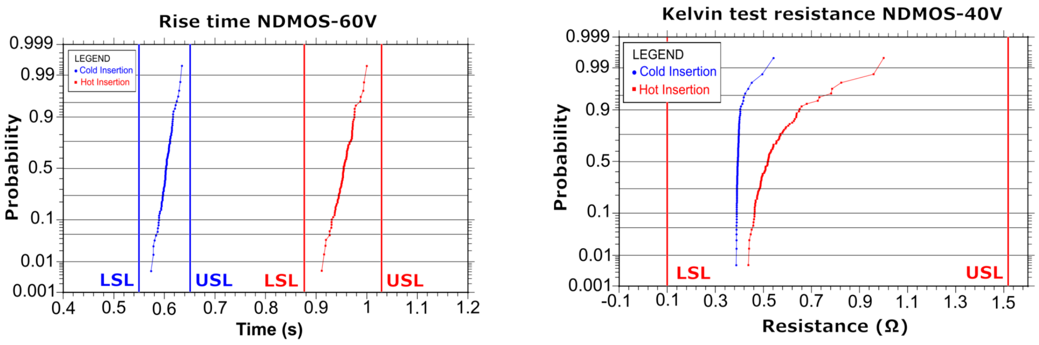


Figure 30: Rise time of the NDMOS-60V and the kelvin test resistance of the NDMOS-40V for both insertions. While the left plot show normally distributed data, the data on the right is tailing towards higher values.

On resistance NMOS-1.5V: Another parametric test is the on resistance measurement of the NMOS-1.5V, for which the data is shown in figure 31. Clearly visible in the left plot, during the hot insertion, is the outlier which gets successfully screened by the USL. The limits have been calculated with a long term C_{pK} of 1.5 once more.

During the cold insertion the outlier is not visible anymore, since it was screened during the hot insertion. Since this is also data from a parametric test, which is affected by temperature differences, the hot and cold insertions have different limits again. The reason for the shift to lower resistances, due to lower temperature is also a known effect already described in chapter 3.4.1.

Apart from the outlier during the hot insertion, the data is normally distributed and does not have a high skewness or kurtosis. At this point it is to mention, that since these were the first devices that have been tested, in order to check the TP and to introduce the first set of limits. There was no root cause investigation of the one outlier device. Why this particular device failed the test is therefore not known.

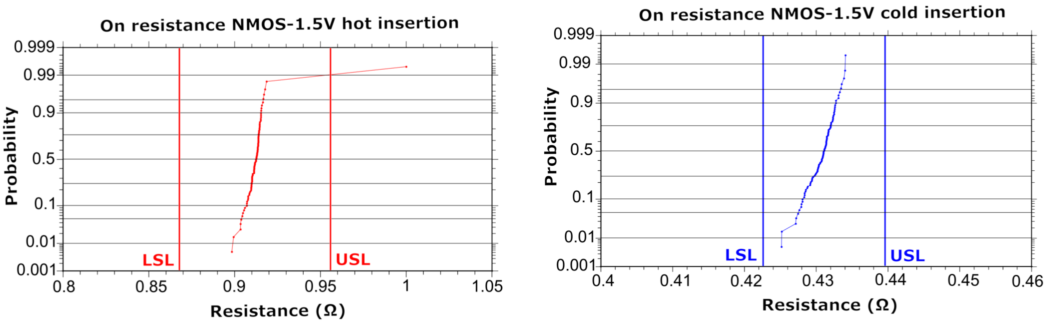


Figure 31: On resistance measurement of the NMOS-1.5V for both insertions. Clearly visible in the left plot is the outlier which gets screened by the USL.

Several additional measurement data for other tests and its distributions are depicted in the appendix. Showing them in this chapter, however, would not help to understand the method of electrical data evaluation and are therefore not included here.

Now that the method of electrical data evaluation has been introduced and verified at actual test data, the next chapter focuses on the burn-in process and the evaluation of the data that is gathered throughout it.

6 Burn-in process and data analysis

As mentioned several times throughout the thesis already, burn-in is a process during which the investigated devices are stressed by high temperature (150 °C) for several hours, even up to days, in order to simulate an ageing process. This ageing has the purpose of activating failure mechanisms that would usually only take effect later in the lifetime of the product. Some of those failure mechanisms and a description of the ELFR can be found in chapter 4. The following pages will focus on the data that was acquired by burning 50 BIRD chips for 48 hours and calculating the resulting drift of parameter values of the tests introduced in chapter 5. This drift will give information on whether any fails have been activated or not.

6.1 Burn-in process

The testflow as depicted in figure 25, is not the complete testflow which every BIRD experiences. Additionally to it, the burn-in and a post burn-in electrical measurement are included after the initial testing.

The full testflow of the BIRD is depicted in figure 32. As can be seen, the burn-in is followed by a hot and a cold insertion. That way, two sets of BE data are gathered, one pre burn-in dataset, for initial screening and one post burn-in dataset.

Bluntly said, the main goal of the burn-in process is to get rid of itself. That is because once enough chips have gone through the burn-in process and no failure mechanisms get activated, the time, each chip spends in the burn-in oven, can be reduced. This is called the burn-in reduction, which eventually leads to fully skip the burn-in process all together. The technology is then considered burn-in free.

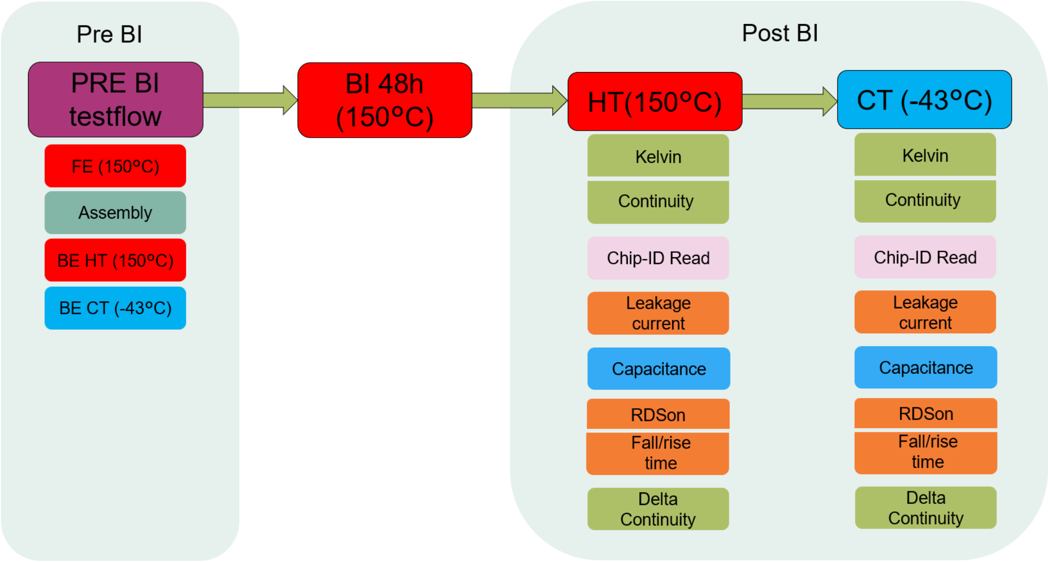


Figure 32: Full testflow of the BIRD, before the burn-in time reduction.

If a technology is burn-in free, it means that the manufacturing process of the devices is optimized to such an extent, that extrinsic failure mechanisms, which can get activated by over-temperature, are so rare, that reliability is assured, even without burn-in.

Even though the devices do not need to undergo the thermal burn-in anymore, they are still being stressed with over-voltage at the stress tests during FE. Since these test are also included in order to activate extrinsic failure mechanisms, they contribute to the reduction of the ELFR and to assure reliability. The difference, however, is that those stress tests can be done in a few milliseconds, in contrast to the burn-in, which initially takes two days.

During the whole BIRD burn-in study, the goal is to show that the technology can be burn-in free, because the stressing with over-voltage is enough to screen the devices with extrinsic failure mechanisms.

Another aspect of the BIRD burn-in study is to see if the methodology is working. Until this point, the technology, which is used for building the chips that are sold, are undergoing the burn-in when they are implemented in the product already. That means that the burn-in study is, at this point in time, something that is being conducted as part of the development process of a new product, whenever new technology is used. The BIRD method is a different one. Since it is not a product, but a chip purely designed for conducting burn-in studies, hopes are that in the future, every technology is reaching burn-in free status though a BIRD. This has the potential to save a lot of money, because of the large reusability of the TP and the data analysis methods as well as the knowledge that was gathered during this BIRD burn-in study.

6.2 Burn-in data evaluation

As established, there are one insertion of FE and several insertions of BE data available for the electrical characterization of the devices. When speaking about post burn-in data, the data from the BE insertions after the burn-in is meant. The difference between the post and the pre burn-in data is called the drift.

The drift is calculated from two sets of measured electrical data by the following formula:

$$drift = \frac{post - pre}{pre} \cdot 100 \tag{48}$$

This is the relative drift in percentage, which gets calculated for each test. *Post* hereby stands for the data gathered during the measurements after the burn-in and *pre* for the data gather before the burn-in. Since there is a chip-ID, the drift can be calculated for each single chip. In the ideal case, when there are no defects in the devices, the measured parameters yield the exact same results after the BI, as they did before the BI. When calculating the difference of those, the ideal drift value is zero for each device and test.

When plotting the data in the same way as explained in chapter 5.2, the resulting distribution looks like a straight, vertical line at zero. In real measurement data, there are slight deviations from the ideal case, caused by noise during measurements for example. This results in drift data that, in the best case is normally distributed around zero, with only a small standard deviation (this is from now on called 'good' drift data). Any deviations from this best case needs to be assessed, whether it is a problem or not.

The advantage of this method is that measurement data of different parameters can be evaluated by one set of common limits. This makes the evaluation faster and easier to understand for people who also might not have the most experience with statistical data evaluation.

One common deviation from good drift data, is that a systematic drift of the whole distribution has occurred due to the burn-in. If this drift is within ±20 %, it is acceptable. If it is outside those limits, it needs to be assessed via a root cause analysis and if the tested parameter is supposed to drift that much.

In the case of a single outlier device, violating the limits, it is likely that the BI activated an intrinsic failure mechanism and the device gets screened from further testing. These devices will be analysed in more detail afterwards, in order to assess what failure mechanism was triggered.

Another common deviation is due to the way, the burn-in data is calculated. When the initial value is closer to zero than the measurement uncertainty and the post burn-in data shifted a little, the drift in percent can be very high, even though the pre burn-in data lies within the limits. This downside to calculating the relative drift in percent is well known and needs to be taken into consideration, when evaluating the data, by looking at the absolute values of the measurement data.

In figure 33 the burn-in data of the rise time of the NDMOS-60V can be seen. It is an example of low drift, well distributed data. The mean value of the drift of the hot insertion was $(-0.45 \pm 0.02)\,\%$ and the mean value of the drift of the cold insertion was $(0.71 \pm 0.02)\,\%$, where the uncertainty is given by the standard error of the mean (SEM).

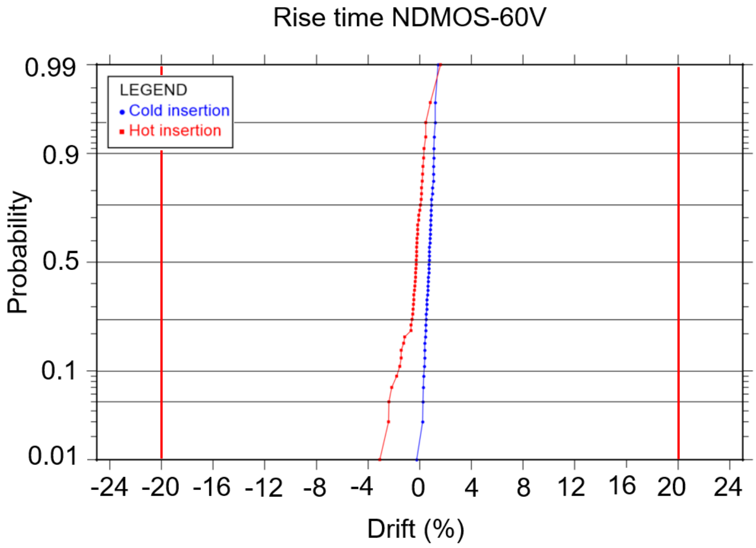


Figure 33: Drift of the rise time of the NDMOS-60V between before and after the 48h BI in %.

Many burn-in data, especially from the parametric tests, look like this, which is why, depicting all those similar looking distributions is refrained from. Instead, table 5 shows the mean drift of each test group, together with their standard error and the number of fails that are occurring due to violating the ±20 % drift limits.

Insertion	Testname	Average drift (%)	SEM (%)	Limit violations
Hot (150 °C)	Kelvin	1.6	7.9	278
	Continuity	2.9	0.3	2
	Leakage current	-12.6	1.0	120
	On resistance	-1.5	0.9	0
	Fall time/rise time	-1.0	0.5	0
Cold (-43 °C)	Kelvin	-13.6	9.0	678
	Continuity	0.3	0.1	0
	Leakage current	-5.9	4.3	133
	On resistance	1.3	2.9	0
	Fall time/rise time	-0.3	0.2	0

Table 5: Average burn-in drift for the relevant test groups, together with their standard error of the mean and the number of drift limit violations during hot and cold insertions.

Not all tests have been included in table 5. Besides tests like the chip-ID and the stress tests, for which a burn-in drift is not possible/meaningful to calculate, also the delta continuity tests are not included. This is because the data used for the delta continuity tests are already the difference between two continuity tests itself. Calculating the drift of data, which for pre as well as post burn-in, lie close to zero, easily results in very high drifts. These drift values, like mentioned above, are not representative of bad burn-in results. They can be from measurement uncertainties or noise and are well within the drift limits when viewed in absolute values instead of percentages.

The results in table 5 are depicted as boxplots in figure 35. As already known, the Kelvin measurements are contact tests which are expected to have different values for every time a contact is made. The limit violations, as well as the drifts of these tests are not meaningful for the assessment of burn-in failures, since as long as the post burn-in contact resistance is below the clamping limit, the test passed, regardless to what the value was before the burn-in. Because of this, the Kelvin test data was not included in the boxplot.

In figure 34 a single boxplot is depicted. The main elements to a boxplot are the box in the middle, which contains 50 % of the data. The line inside the box represents the median, with left and right to it, being 25 % of the data each. On both sides, outside the box, are the so called antennas. They reach to the highest value inside a $1.5 \cdot$ IQR interval subtracted from the 25 % value and added to the 75 % value. IQR stands for interquantile range and is the value range between the upper and lower box limit values.

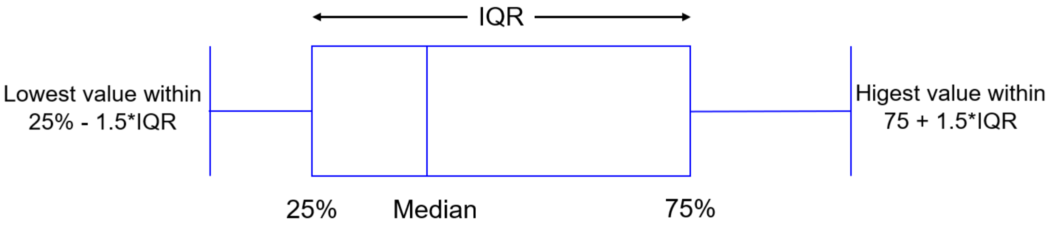


Figure 34: Demonstrative boxplot.

In figure 35, one can clearly see that the continuity, on resistance and fall/rise time have a low mean drift of below ± 5 %, while the leakage current has a higher mean drift as well as a greater range. Especially the data from the cold insertion oversteps the -20 % limit and has by far the greatest range of all the tests.

The reason for this lies again in the low values of the leakage current. Because the leakage current is especially low at low temperatures, the measured values are often times close to zero in the pre burn-in insertion. The shift is within the measurement uncertainty and needs to be looked at in absolute values.

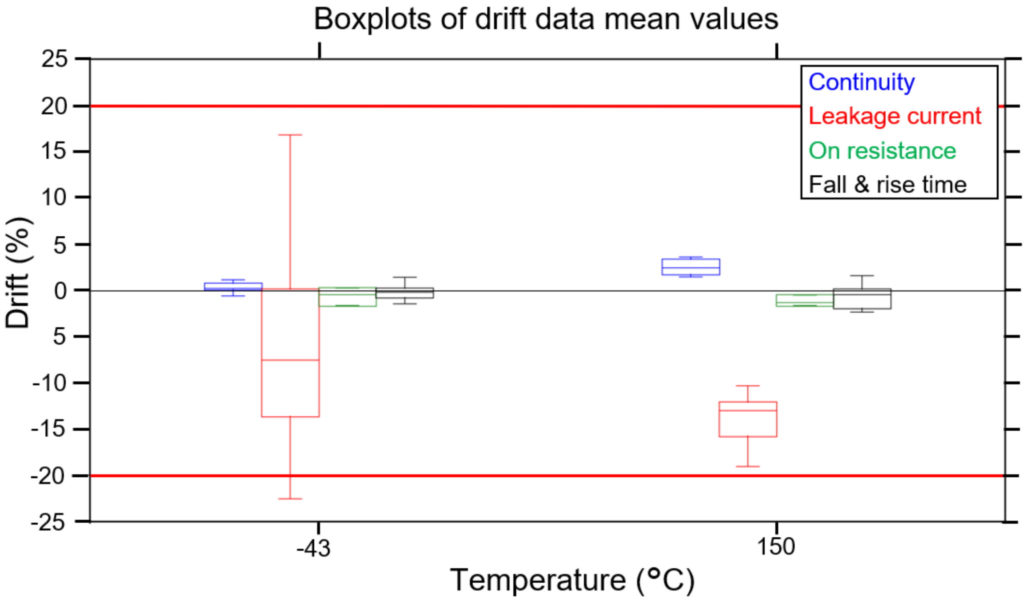


Figure 35: Boxplots for the mean values of each test in the categories in table 5 (without Kelvin) for the hot and cold insertion.

Due to the mentioned measurement uncertainty and the general setup dependency of the leakage current tests, all the limit violations in table 5 can be explained. These tests also have large standard deviations in comparison to other parametric tests. An actual activated latent defect would increase the absolute value of the leakage current, by orders of magnitude. The resulting outlier would not only be caught by the drift limits, but also by the limits of the measurement data. This was never the case in the data gathered here.

A representative plot for the leakage current drift data can be seen in figure 36. On the left side is the pre and post burn-in data distribution for the cold insertion, whereas on the right side the respective relative drift of those two is depicted. The leakage current after the burn-in is decreasing for nearly every device according to this data, which is highly unlikely. Much more likely is that due to the very low currents, measurement uncertainty and setup differences are primarily responsible for the high drift and the decrease in leakage current.

It is clearly visible, that even though there are drifts of nearly +20 % and -65 %, no failure was activated during BI, because the post burn-in data do not show any outliers. This is the case for all other drift limit violations during the leakage current tests as well.

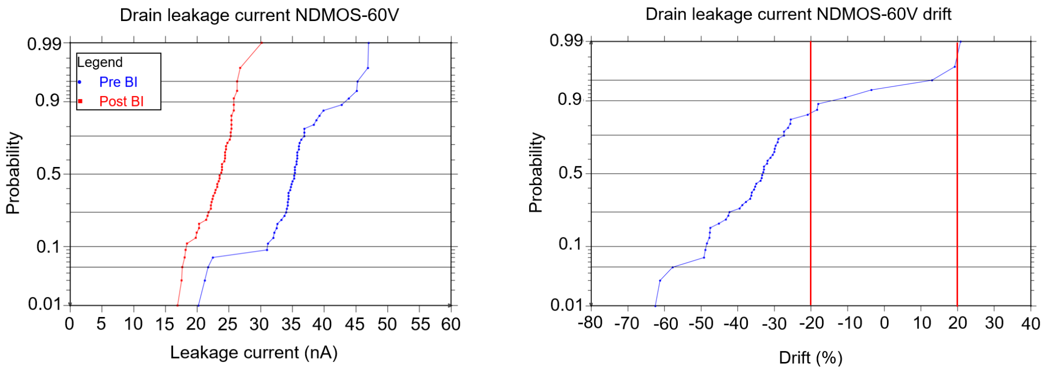


Figure 36: Drain leakage current test data of the NDMOS-60V during the cold insertion before and after 48h burn-in on the left and the resulting drift in percent on the right.

Regarding the rest of the parametric tests, there were no limit violations and the mean drifts, as well as their standard deviations were in a low range, as was shown in figure 35.

After assessing all the burn-in data for 50 devices and examining every distribution carefully, no valid fails were found. The technology is therefore showing promising first results. Since, however, this thesis aims to assess whether real device defects can be found with the help of the TP and the burn-in methodology, this promising result is not enough. The next chapter focuses on the limitation of the devices and under which circumstances they would have failed. In this final chapter the goal is to simulate conditions under which a valid failure would have occurred and to what extent the electrical test result would have differed from the data shown so far.

7 Burn-in failure root cause identification

Since there were no real burn-in failures found in the 50 BIRD chips that were stressed and tested, several of them were modified in order to simulate conditions that would have resulted in a failure at some point during the burn-in. This can be done in the failure analysis (FA) laboratory. Most of the times the FA laboratory is used to analyse devices which have already been sent to the customer, that have failed or malfunctioned during field application or to analyse devices that failed during some kind of verification process (like the burn-in). It can, however, also be used to modify good devices, such that they show different characteristics than before. A couple of such modifications were thought of, however, not all of them proved to be working. All of them included a focused ion beam (FIB) modification at or under the surface of the chip, which made it necessary to remove the top of the package and the protective top layer of the chip first.

7.1 Gate oxide breakdown

The first goal was to stress the devices on the BIRD with such a high voltage, that the gate oxide breaks down. However, applying a large overvoltage to the transistors proved to be quite difficult, because of the ESD structures and all of the inverter structures on the gate and on the enable line, which shield the device under test from those voltage spikes. Therefore, a modification was done on the NDMOS-60V using a FIB.

7.1.1 Focused ion beam (FIB)

FIB is a general method of cutting through or depositing small structures with the means of a (mostly) metallic ion source. This method is not only used in the semiconductor industry and research, but also, among others, in material science, biology and geology.

A usual FIB instrument consists of a vacuum system, a liquid metal ion source, an ion column, some detectors and a gas delivery system. The vacuum system provides vacuum conditions around the sample and the ion source/column in order to avoid particles or gas molecules to decrease the mean free path of the ions and scatter them. The liquid metal ion source provides the ions, which are used to modify the sample. As the name indicates, metal is molten and the liquid then coats the ion column, which is usually made out of tungsten. Often times the liquid metal is gallium, but it can also be other metals. At the tip of the ion column, the liquid metal then forms a sharp cone (called a Taylor cone) due to the applied electric field between the ion column and an extraction electrode. The ions are released from this tip by the electric field and get condensed by a system of electromagnetic lenses.

In order to deposit different metals on the surface of the sample, a gas delivery system can introduce a tungsten or platinum gas for example. Through ion beam induced deposition (IBID), the organometallic precursor gas (the gas containing the metal) gets decomposed only where the ion beam hits the surface and the metal deposits.

It is also not unusual to include a scanning electron microscope (SEM) or other detectors, like secondary electron detectors, in the same instrument, in order to take images during and after the FIB modification.

Figure 37 shows the principle of the liquid metal ion source, forming the Taylor cone on the tip of the tungsten ion column.

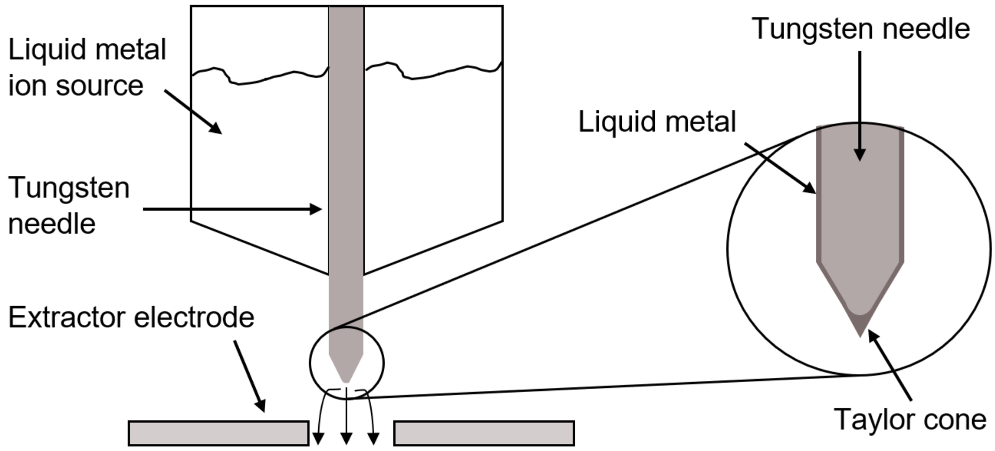


Figure 37: Schematics of the ion beam generation system of a FIB setup.

This was only a brief summary of the working principle of a FIB setup. For more detailed information see [33].

In the scope of this thesis two techniques were used to prepare the samples. Once, a FIB was used to make a precise cut through layers of the chip, which makes it possible to sever structures from each other or to examine cross sections of vertical structures. The other technique is the ion-beam-induced deposition (IBID) mentioned above.

7.1.2 Sample modification

In order to sever the connection between the gate contact of the NDMOS-60V and the gate/enable line, which leads to the contact pins of the chip, a FIB cut was made. The cut severed the metal lines connecting the ESD structure and inverters from the gate contact of the transistor. A cross section of the modification can be seen in figure 38 with the mentioned FIB cut being the trench on the left side, closer to the gate/enable pin.

Since the contact pins were now disconnected from the gate as well, a new contact needed to be placed. To realize this, another FIB cut was made, this time, however, closer to the gate contact, above the metal line, which was formerly used to connect the gate to the pins. By cutting down into the metal and filling the hole up with tungsten, a new vertical connection to the surface of the chip was established. Afterwards, a tungsten cross was placed on the surface of the chip, in order to better contact the gate from above. The filling up of the hole and the placing of the cross was done using IBID.

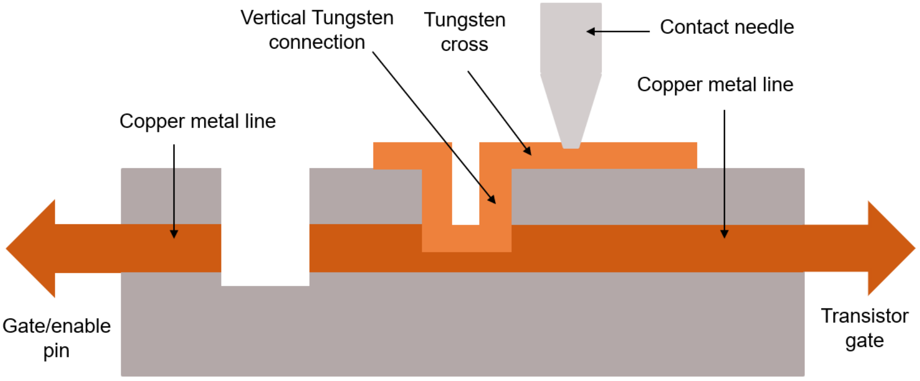


Figure 38: Cross section of the FIB modification for the oxide breakdown. Two FIB cuts were made and one of them was filled up with tungsten, with an additional measuring cross on top, in order to better contact it.

Figure 39 shows the initial FIB cut, together with the second cut and the tungsten cross.

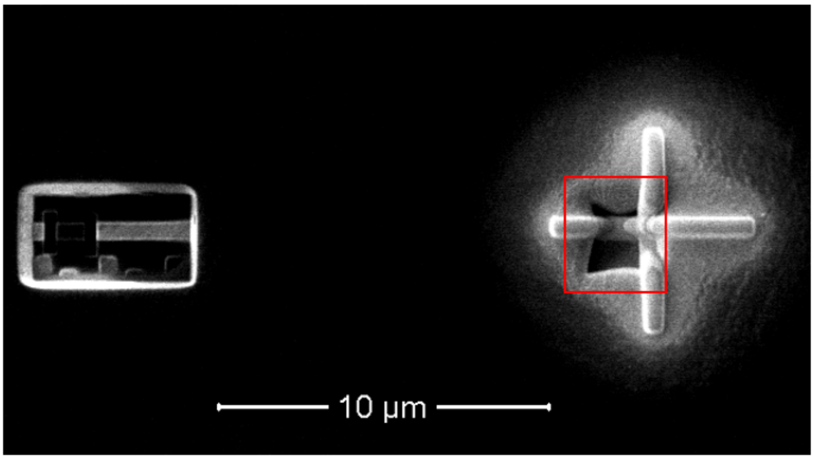


Figure 39: SEM image of the FIB cut on the left, with the tungsten cross on the right. The second FIB cut, which contains the vertical connection to the copper metal line, is visible inside the red box.

Afterwards, the Tungsten cross was contacted by a probing needle, while the drain contact and ground (which is connected to every source contact of each transistor on the BIRD) were contacted via the measuring board.

A picture of this setup is included in figure 40. Above the prober table, which contained the board with the device under test in it, is a microscope, in order to adjust the needle and the device properly.

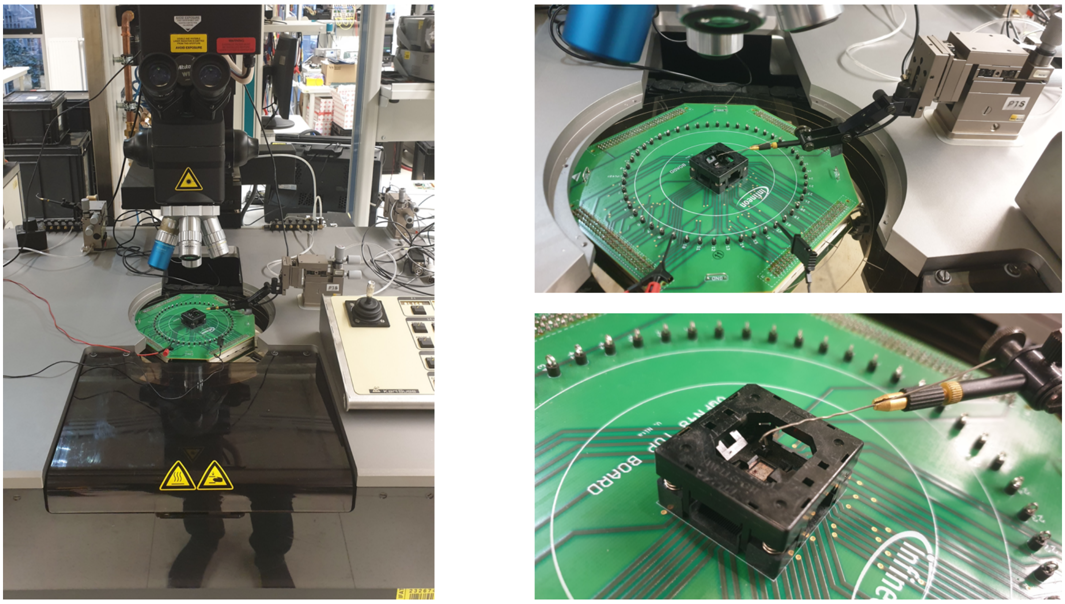


Figure 40: Setup of the prober table, used for contacting the measuring cross with the micro needle, in order to measure the gate oxide breakdown of the NDMOS-60V. On the left the whole probing station is visible, with the microscope, and the probing table. On the top right a closeup of the (green) measuring board is depicted and on the bottom right, a close up of the (black) chip socket for the VQFN-48 package and the probing needle is visible.

7.1.3 Measurement procedure

The first measurement performed on the device was to ramp the gate voltage and measure the drain current, until the saturation current was met. Afterwards, the gate and the drain leakage were measured. The third measurement included to ramp the gate voltage even further, until the oxide breakdown. After the breakdown, the gate and drain leakage were measured again.

Several measurements at ambient temperature were performed on this device, in the following order:

1. The drain voltage on pin 9 was set to 0.05 V. The gate voltage, contacted with the needle was ramped from 0.0 V to 2.5 V and the drain current was measured.

2. The drain leakage current was measured, while the gate, contacted with the needle, was biased with 5.0 V and drain on pin 9 was at 0.0 V. Afterwards the gate leakage current was measured, while the gate, contacted with the needle, was unbiased and the drain voltage on pin 9 was 2.5 V.

3. The drain voltage on pin 9 was set to 0.05 V. The gate voltage, contacted with the needle, was ramped from 2.5 V until the oxide breakdown, while both, the drain and the gate current were measured.

4. Measurement 2 was repeated.

7.1.4 Measurement results and discussion

The measurement results from the drain current measurement can be seen in figure 41. It is visible that a good contact was established with the needle and that the measured drain current as a function of the gate to source voltage is as described in chapter 4.3.2. and figure 22. The subthreshold swing and the extrapolated threshold voltage are magnified in the plot on the bottom right.

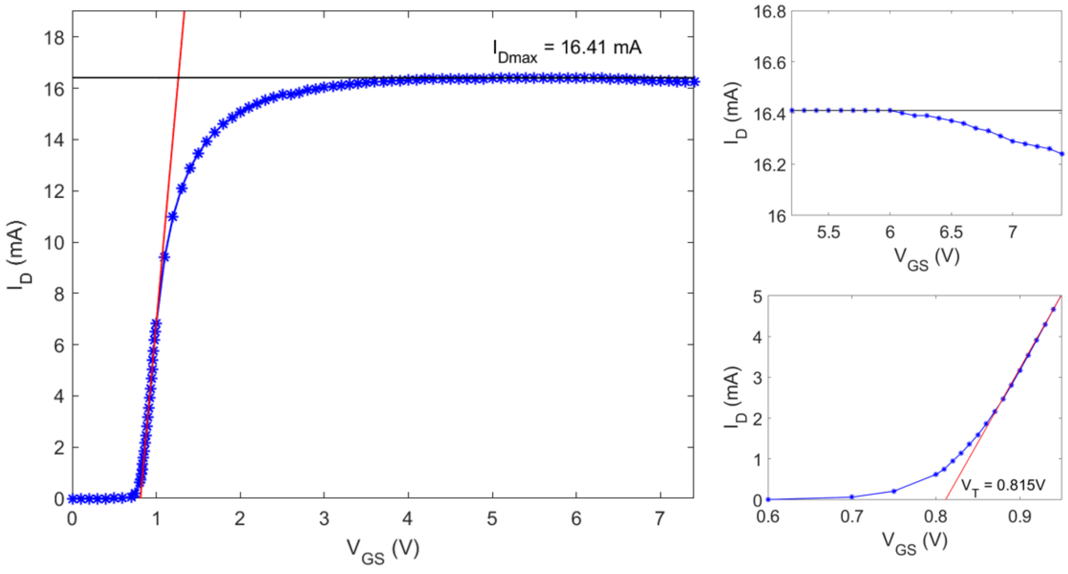


Figure 41: Drain current data of the NDMOS-60V as a function of the gate to source voltage. A zoom to the subthreshold region is included on the bottom right, with the red line being a linear fit of the data in the region where the gradual channel approximation is valid. On the top right a zoom into the maximum voltage region, where the drain current starts dropping is shown.

The threshold voltage for this device lies at $V_T = 0.815\,\mathrm{V}$. The maximum drain current is reached at 5.2 V and stays constant at $I_{D_{max}} = 16.41\,\mathrm{mA}$ until 6.0 V are reached. The inversion channel is in pinch off in that state, which means that the additional voltage is increasing the electric field, but not the drain current anymore. After that, the drain current starts to slowly decrease. This can be explained by the self heating of the device during operation. As mentioned in chapter 3.4.1, the on resistance rises with increasing temperature and therefore the drain current decreases. The slight decrease in drain current can be seen in figure 41 on the top right.

At this point the leakage currents at the source and drain contact with the mentioned conditions above were measured. The results of this measurement can be seen in table 6.

Next, the gate current was measured, as a function of the gate voltage. Even though the gate current should be zero, because of the insulating gate oxide, there is still a remaining gate leakage current, that is caused by the insulator conduction mechanisms, introduced in chapter 4.3.1.

Several insulator conduction mechanisms can be ruled out as the leading one

measurement	$I_{D_{leak}}$	$I_{G_{leak}}$
pre breakdown	2.8 nA	0.1 nA
post breakdown	100 μA	100 μA

Table 6: Leakage current measurements before and after the oxide breakdown. The current measured at the post breakdown measurement was the compliance current. No higher current was measured, in order to protect the prober needle.

from the start, due to the thickness of the gate oxide and the temperature range at which the measurement was done. Because direct tunneling is negligible for gate oxides thicker than 3.5 nm, which is true for the 1 tu thick GOX3, (see table 1), it can already be ruled out. Schottky emission can be ruled out as well, since the increase in insulator current was strongly depending on the electric field and Schottky emission is dominated by thermal activation. Therefore also thermionic-field emission can be neglected, since it also requires higher thermal energies. Because of the exponential nature of the increase in gate leakage current, ohmic conduction can be ruled out, since this mechanism has a linear dependence on the electric field.

The remaining conduction mechanisms are the F-N tunneling, F-P emission and hopping conduction, for which the formulas are given below:

$$J_{FN} = \frac{q^2V^2}{16\pi^2\hbar\phi_B d^2} exp\left[\frac{-4\sqrt{2m_T^*}d}{3\hbar qV}(q\phi_B)^{3/2}\right] \tag{49}$$

$$J_{FP} = q\mu N_C \frac{V}{d} exp\left[\frac{-q(\phi_T - \sqrt{qV/d\pi\epsilon_i\epsilon_0}}{k_BT}\right] \tag{50}$$

$$J_{hopp} = qanvexp\left[\frac{qaV}{dk_BT} - \frac{E_a}{k_BT}\right] \tag{51}$$

In those the current density J is given in terms of the electron charge q, the electric field V, the thickness of the oxide d, the Schottky barrier height ϕ_B, the tunneling effective mass m_T^*, Planks constant h, the electron mobility in the insulator μ, the density of states in the conduction band N_C, the trap energy level ϕ_T, the temperature T, Boltzmann's constant k_B, the mean hopping distance a, the electron concentration in the conduction band of the insulator n, the frequency of thermal vibration of electrons at trap sites v and the activation energy of the traps E_a.

The measurement data of the gate current as a function of the gate voltage can be seen in figure 42.

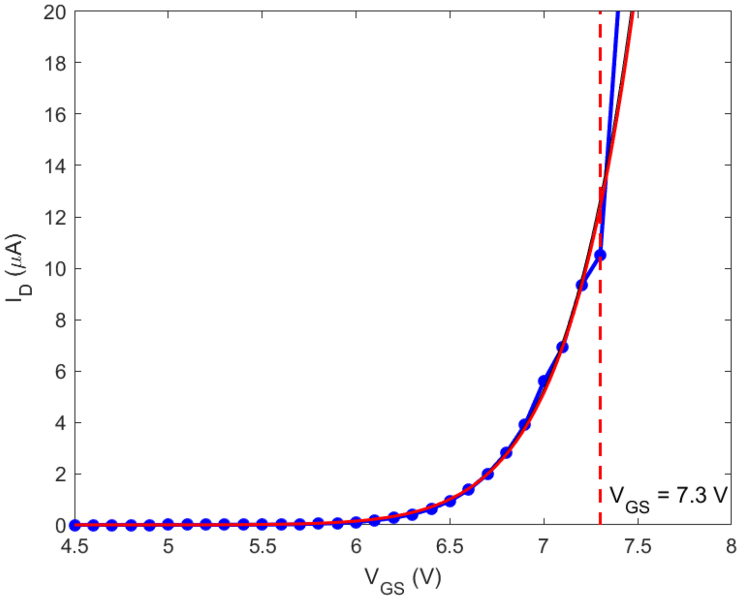


Figure 42: Gate leakage current measurement data as a function of gate voltage. The red line is a fit according to equation 49, which is the theoretical expression for the F-N tunneling.

When trying to fit the measured data with the three formulas for the conduction mechanisms, the theoretical curve for the F-N tunneling fits the data. The values, used for the Schottky barrier height ϕ_B was 3.1 eV and the value for the effective tunneling mass m_T^* was taken as $0.5 \cdot m_0$, after [34].

Trying to fit the measurement data with the theoretical curves of the F-P emission and hopping conduction, proved to be difficult, since several parameters, like ϕ_T, a, v and E_a were unknown and could have only been determined through additional experiments which could only be conducted through the use of various other experiments. Using different literature sources, of determining these parameters would have created great uncertainties. Because of this circumstance and because the data is already well fit by the F-N tunneling curve, it was refrained from fitting the remaining two conduction mechanisms to the measurement data and F-N tunneling was accepted to be the leading conduction mechanism.

For further thoughts on how to determine the unknown parameters in equation (50) and (51), see chapter 8.

The deviation of the measurement data from the theoretical F-N curve at high voltages is due to the breakdown of the oxide. Above 7.3 V the leakage current is not considered to be due to F-N tunneling anymore, but rather due to a short of the gate oxide. A conduction path through the insulator was established, which causes the gate current to not follow the theoretical path of F-N tunneling anymore.

Measurement four then yielded the following result: $I_{D_{leak}} = 100\,\mu\text{A}$ at the drain contact and $I_{G_{leak}} = 100\,\mu\text{A}$. It is assumed that the actual value of the drain and gate leakage currents is even higher, however, in order to protect the thin needle of the prober from melting, due to too high current densities, compliance was set to $100\,\mu A$. Table 6 shows these results together with the results of measurement 2.

Comparing the gate leakage current from before the oxide breakdown, to after the oxide breakdown, an increase of $1 \cdot 10^6$ was observed, while for the drain leakage current an increase of around $3 \cdot 10^4$ was observed. The left hand side of figure 43 shows the gate leakage test data for the cold insertion before ($\mu \approx 5.5\,nm$) and after ($\mu \approx 7.0\,nm$) the burn-in for the NDMOS-60V, where μ is the mean value of the distributions in this case. All measurement values are below 10 nm during this measurement. Comparing the values for the post burn-in measurement, to the values of the $I_{G_{leak}}$ for the pre breakdown measurement, one can see that the values, are about a factor of 70 higher. This is explained by the parasitic leakage currents from the measurement equipment and the circuit of the enable line, which were not present in the measurement of table 6.

When comparing the post burn-in data of the left plot in figure 43 to the post breakdown $I_{G_{leak}}$ measurement, however, one can see that the increase in gate leakage current is still in the order of $1.5 \cdot 10^4$. This big increase in gate leakage current, upon gate oxide breakdown, is clearly visible in the data distribution as well. The right hand side of figure 43 shows the same data as the left hand side, except for one data point, where the post burn-in measurement was substituted with the post breakdown gate leakage current of the device modified with the FIB.

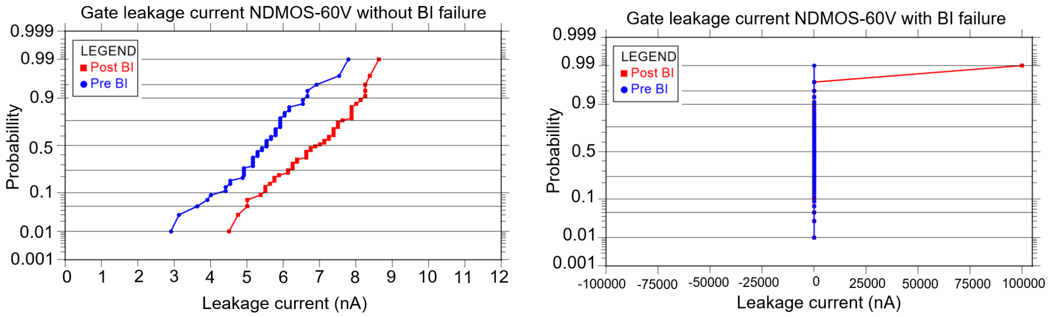


Figure 43: Gate leakage current test data of the NDMOS-60V during the cold insertion before and after 48h burn-in on the left side. On the right side, the same data was plotted, except that the post burn-in drain leakage data of one device was substituted with the post breakdown $I_{G_{leak}}$ data from table 6.

Similar looking plots would result when substituting the post breakdown drain leakage current with its respective data point in figure 36. It is clearly visible that when the gate oxide breaks down, after the burn-in, the device(s) have a leakage current, which is of orders of magnitudes higher than the intact ones. This result reassures, that during the burn-in of the 50 devices, which were shown in chapter 6, no valid burn-in failures have been observed, even though, several tests showed relative drift data which were outside the limits. Furthermore, it shows that the $C_{pK} = 1.5$ limits, which were introduced in chapter 5.3, detect and screen out such failures as well.

7.2 Source to drain short

In order to introduce a second mechanism which can be activated during the burn-in, a short between the drain and the source has been decided upon. This can happen, during the fabrication process of the copper metallization. The most used process for contacting the transistors and fabricating the copper connections between them, is called the dual-damascene process. [35]

Dual-damascene process: In its most basic form, the dual-damascene process consists of four process steps. During the first step, trenches are being etched inside of the silicon dioxide, through the use of photoresists, as described in chapter 4.3.3. These trenches then get coated with a diffusion barrier, which serves the purpose of preventing the copper to diffuse into the oxide. Afterwards, the copper gets deposited into the trenches, followed by a chemical-mechanical planarization (CMP), where the surface gets polished and the excess copper is removed. If the etching of the trenches, for example, is not optimized properly, it can happen that the two metals which should not be connected, are connected by the deposited copper. Additionally to that, residues from the CMP can leave traces of copper between the metal lines, which can result in line to line leakage or shorts. This can happen, between the source and the drain contact of the MOSFET for example.

Upon stressing the device, with a high drain voltage, this parasitic, ohmic connection can be destroyed, because the high current density melts the metal. Before stressing the device with the overvoltage, the gate leakage current is therefore high, however, decreases at the second leakage current measurement, after the stress voltage was applied (see chapter 5.1.1). When calculating the delta of the pre and post overvoltage stress leakage current measurements, this decrease can be seen and screened by the test limits.

Since no such failure was observed in the 50 BIRD chips that were stressed, this failure had to be artificially introduced as well.

7.2.1 Sample modification

For this experiment, three different BIRD chips have been modified with the FIB. The technique which was used, was the IBID, which was also used to deposit the tungsten cross in the previous modifications. A connection between the source and the drain contact of the NMOS-1.5V was introduced, using platinum. Figure 44 shows a SEM image of one of these modifications.

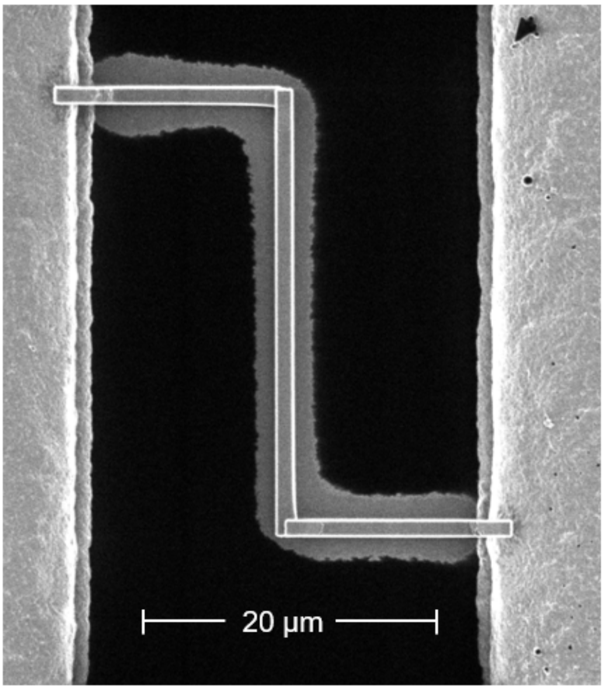


Figure 44: SEM image of the IBID modification of one of the samples. Platinum was deposited in order to establish a conductive connection between the source (left copper) and the drain (right copper).

By depositing different amounts of platinum, different resistances can be achieved. The dimensions of the depositions were chosen such, that the introduced resistance would cause a leakage current close to the upper test limit of the drain leakage current tests. The devices that were used for this, were initially at the lower end of the distribution, because the goal was to see a maximum drift of the drain leakage current between before and after the platinum connection was molten.

samplename	length (μm)	width (μm)	height (μm)	resistance (Ω)
smpl 1	60	1	1	1200
smpl 2	60	0.8	0.7	2140
smpl 3	60	0.8	0.7	2140

Table 7: Dimensions of the deposited Platinum together with the resistances that wanted to be achieved.

In table 7 the chosen dimensions of the deposited platinum lines can be seen, as well as the resistances which were theoretically achieved, due to the dimensions of the deposition.

7.2.2 Measurement procedure

The measurements that were performed were the same for all the devices:

1. The drain voltage on pin 23 was increased manually from 0 V until the measured current at the drain contact showed a significant drop, while gate (pin 24) and enable (pin 25) were kept at 0 V. Source was connected to the global ground contact of the chip.

2. The drain voltage was increased manually from 0 V again, after the current drop had occurred, until the maximum voltage of the first measurement, whilst the drain current was measured again. Gate, enable and source were the same as before.

7.2.3 Measurement results and discussion

The measurement data of all three devices is shown in figure 45.

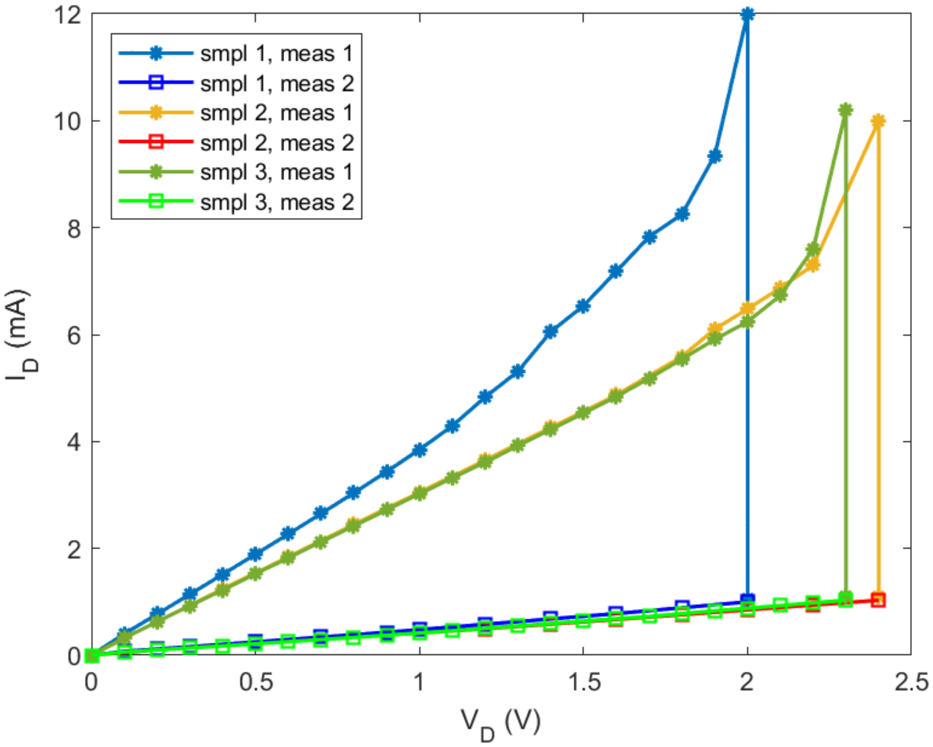


Figure 45: Drain current measurement as a function of drain voltage for the three modified samples. The point shaped data points are the first measurement, while the square shaped data points are the second measurement, after the current drop occurred.

The first thing that can be seen, is the high drain leakage current that was measured for all three samples. For the first sample, the maximum leakage current was 12 mA, which was reached at 2 V. Before the drain voltage could have been increased again, the current dropped significantly to 1.06 mA. The maximum current of the second measurement on smpl 1 was then only 1.06 mA at 2.0 V.

Before and after the measurements, microscope pictures of the modification were taken at the probing station. The goal was to visually inspect them for eventual burn marks, after the current drop occurred. Figure 46 shows the modification which was shown in figure 44, but through an optical microscope. On the left hand side is a picture from before the measurements, while on the right hand side is the same modification after the measurements. One can see that the modification melted in the area which was highlighted by the red box. The current drop, which occurred on all three samples, was due to this melting.

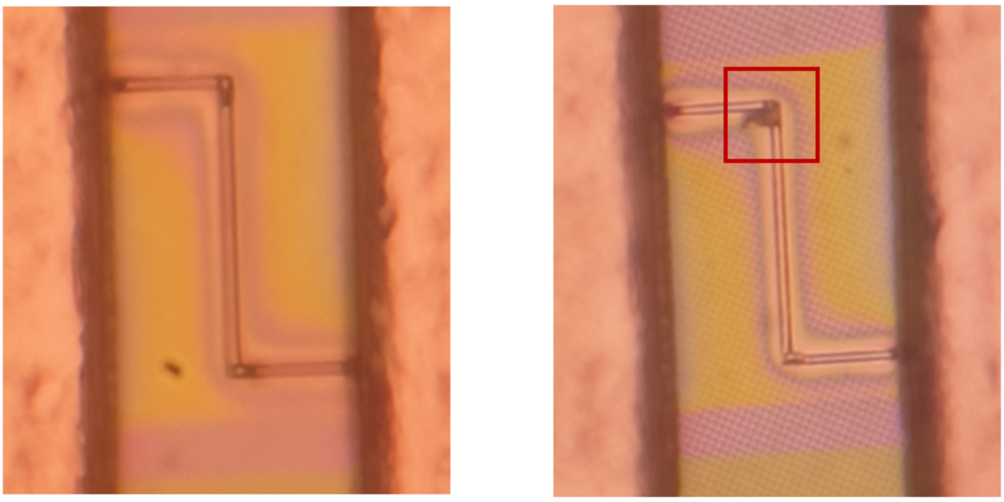


Figure 46: Picture through an optical microscope of the source drain short FIB modification of the NMOS-1.5V before (left) and after (right) the leakage current measurements. Inside the red box, one can see the part of the platinum connection which melted due to the high current densities.

Another thing that is visible, is that the maximum drain voltage that was measured is below 2.5 V, which is less than half of the operating drain voltage for this device (5.0 V).

The fact that relatively high currents were reached at not even half of the operating voltage, indicates, that the desired resistance was not reached during the deposition process. From the plotted graphs, one can see the initial ohmic increase in drain current with voltage, which is expected, since the introduced platinum acts as a parasitic ohmic resistance. At higher voltages the current starts to increase even faster, which can be explained by the melting of the metal. The slope of the linear range of the measured data is one over the actual resistance which the current experienced, as according to Ohm's law. Table 8 shows the highest measured drain current at both measurements, together with the respective drain voltage, as well as the calculated resistance prior to the current drop, for each sample.

As suspected, the resistances of the deposited platinum lines were not nearly as

sample name	$I_{D_{max1}}$ (mA)	$I_{D_{max2}}$ (mA)	$V_{D_{max}}$ (V)	R (Ω)
smpl 1	12.0	1.06	2.0	260
smpl 2	10.0	1.03	2.4	331
smpl 3	10.2	1.03	2.3	329

Table 8: Maximum drain current values for each sample and measurement, together with their respective drain voltage and resulting resistance.

high as intended. This results in a melting of the platinum at voltages below the operating and stress voltage and in drain leakage current values which would be screened out anyways.

One reason which could explain this discrepancy between the expected and the actual resistance values for the platinum deposition is that platinum IBIDs can cause contamination in the vicinity of the main deposition area. This can alter the electrical properties of the deposition, as well as the whole device quite drastically, as can be seen in [36].

Because the actual resistance values did not match the values that were planned, no further investigations regarding this failure were done, since there was also no time and resources left to redo the modifications within the scope of this thesis. For future modifications that intend to introduce a resistance, this way of doing it is not recommended, since the risk of contamination is quite high. Note, that if the deposition had produced the wanted resistances, it would have still been necessary to reference the current capacity of platinum (the maximal current density until melting occurs) to the one of copper, since the CMP residues would have also been copper.

As can be seen in this chapter, not all experiments and modifications were successful, it was, however, successful to prove that no real burn-in failures due to oxide breakdown were observed after the 48h burn-in of the 50 devices. This then, marks the end of the main parts of this master's thesis, with only the conclusion and outlook being left. As the name already indicates, the next and final chapter summarizes the findings of all the previous chapters into a few compact pages and also gives a glimpse into the future proceedings of the BIRD burn-in study.

8 Conclusion

Over the past 7 chapters of this thesis, a burn-in reference device (BIRD) and its purposes has been introduced (see chapter 1 and 2). The goal of the thesis was to show, how with the help of burn-in, early life failure mechanisms could be detected in the new (power) MOSFETs of Infineon Technologies.

To achieve this goal, different kind of failure mechanisms and what influence they have on the performance of the devices and further, their reliability, have been given (see chapter 4.3 - 4.5). It is important to ensure reliability of the manufactured devices, in order to comply with all the industry standards and norms, like the ISO26262, which regulates the functional safety standards for the automotive industry (see chapter 4.1 and 4.2).

The most prominent failure mechanism was the breakdown of the gate oxide, which results in a complete failure of the devices functionality to act as a voltage regulated current switch. Eventually every device will experience such a failure (if no other failure is activated first). However, depending on the quality of the gate oxide, it happens sooner for those with low quality gate oxides or later with high quality ones (see chapter 4.3). In order to screen out the devices which have a low quality gate oxide, the devices are stressed. This stressing is called the burn-in and is realised through increased environmental temperature, while they are being operated (see chapter 6.1).

The increased temperature, similar to the voltage stressing, which is happening during FE testing of the devices, is accelerating the degradation of the devices. 50 BIRD chips underwent the 48 hour long burn-in, in order to investigate if any of the devices experienced a degradation high enough to fail after the burn-in.

To assess, whether a devices fails due to the burn-in or not, all the devices were tested three times before the burn-in and twice after the burn-in. These tests included parametric tests like the on resistance, the fall and rise time or the leakage current of the devices (see chapter 5.1).

By using statistical methods and introducing limits, which screen out malfunctioning devices prior to experiencing the burn-in, it was assured, that all the devices were working properly (see chapter 5.2). After the burn-in, the test data post burn-in was compared to the test data pre burn-in, by again using statistical methods. It was found that no defects were activated, which would have resulted in a failure. This is a great first result, as it shows that according to the sample size of 50 BIRD chips, the manufacturing process is dependable enough to produce reliable devices (see chapter 6.2).

However, in order to be sure, that the theory of the described failure mechanisms was correct and the limits which are used to screen out devices which are effected, by the breakdown of the gate oxide are appropriate, some devices were modified, using a focused ion beam (FIB). The measurement results of the gate oxide breakdown could reproduce the theoretical predictions of the failure mechanism. Fowler-Nordheim tunneling was found to be the leading insulator conduction mechanism, before the breakdown of the gate oxide, due to reaching the critical field happened. The measured leakage current of the device was several orders of magnitude higher than the leakage current of the intact ones and would have been screened out successfully by the limits, if a gate oxide breakdown had occurred during the burn-in (see chapter 7.1).

Another experiment was to introduce a short between source and drain of the devices by using IBID. This ought to simulate a fabrication failure, at which metalic residues remain after the CMP process step, which can lead to an increased leakage current between source and drain. The experiment failed, because the deposited platinum, which acted as an ohmic resistor, did not have the intended resistance. This was probably due to a high contamination around the deposited material, which lowered the resistance, resulting in a high leakage current. Such devices would have been screened out before the burn-in process, without ever activating the defect (see chapter 7.2).

At this point, also a short outlook into what can and will be done next is given in the following sentences. Since the burn-in of the 50 devices that have been tested during this thesis, yielded excellent results, with no devices failing, the whole burn-in study will be conducted next. Over 100 000 BIRD chips will be produced and experience the same testflow as shown in figure 32. If enough devices pass the burn-in without failure, the manufacturing process of the these devices can be regarded as so optimized, that it is not necessary to stress them with a burn-in anymore. Afterwards, the devices will only be stressed by overvoltage during FE testing, which is much faster and therefore cheaper.

List of Figures

List of Tables

References

[1] M. F. Zakaria, Z. A. Kassim, M. P.-L. Ooi, and S. Demidenko, "Reducing burn-in time through high-voltage stress test and weibull statistical analysis," *IEEE Design Test of Computers*, vol. 23, no. 2, pp. 88–98, 2006.

[2] H. Lewitschnig and J. Goncalves, "Design for stress," in *2021 Annual Reliability and Maintainability Symposium (RAMS)*, 2021, pp. 1–5.

[3] S. SZE and K. N. KWONG, *Physics of semiconductor devices.* Wiley-Interscience, 2007.

[4] S. Jasprit, *Semiconductor Devices Basic Principles.* John Wiley Sins, Inc., 2001.

[5] D. Binkley, *Tradeoffs and Optimization in Analog CMOS Design.* John Wiley Sons, 2008.

[6] J. Baliga, *Fundamentals of Power Semiconductor Devices.* Springer, 2019.

[7] S. Saha, *FinFET Devices for VLSI Circuits and Systems.* CRC Press, 2020.

[8] [Online]. Available: https://www.circuitbread.com/ee-faq/what-isa-finfet

[9] T. Erlbacher, *Lateral Power Transistors in Integrated Circuits.* Springer, 2020.

[10] J. Wang, "Mosfets power consumption and power dissipation calculation," *J. Phys.: Conf. Ser.*, vol. 1754, 2021.

[11] J. Appels and H. Vaes, "High voltage thin layer devices (resurf devices)," in *1979 International Electron Devices Meeting*, 1979, pp. 238–241.

[12] A. Ludikhuize, "A review of resurf technology," in *12th International Symposium on Power Semiconductor Devices ICs. Proceedings (Cat. No.00CH37094)*, 2000, pp. 11–18.

[13] J. Fernandez, S. Hidalgo, J. Paredes, F. Berta, J. Rebollo, J. Millan, and F. Serra-Mestres, "An on-resistance closed form for vdmos devices," *IEEE Electron Device Letters*, vol. 10, no. 5, pp. 212–215, 1989.

[14] C. T. Huang, B.-Y. Tsui, H.-J. Liu, and G.-L. Lin, “The impact of high-voltage drift n-well and shallow trench isolation layouts on electrical characteristics of ldmosfets,” in *2007 IEEE Conference on Electron Devices and Solid-State Circuits*, 2007, pp. 267–270.

[15] R. Ye, L. Lu, S. Liu, H. Wu, H. Chen, W. Sun, S. Lu, L. Zhang, W. Wu, W. Su, B. He, F. Lin, and G. Sun, “Reliability concerns on ldmos with different split-sti layout patterns,” *IEEE Transactions on Electron Devices*, vol. 67, no. 1, pp. 185–192, 2020.

[16] J. McPherson, *Reliability Physics and Engineering.* Springer, 2010.

[17] J. Standard, “Method for developing acceleration models for electronic component failure mechanism,” *JEDEC Solid State Technology Association, USA*, 2003.

[18] M. Kimura, “Field and temperature acceleration model for time-dependent dielectric breakdown,” *IEEE TRANSACTION ON ELECTRON DEVICES*, vol. 46, 1999.

[19] T. Prodromakis, P. Georgiou, K. Michelakis, and C. Toumazou, “Effect of mobile ionic-charge on cmos based ion-sensitive field-effect transistors (isfets),” in *2009 IEEE International Symposium on Circuits and Systems*, 2009, pp. 2165–2168.

[20] F. Chiu, “A review on conduction mechanisms in dielectric films,” *Advances in Material Science and Engineering*, 2014.

[21] M. Wang, S. Deng, T. Wang, B. Cheng, and J. Lee, “The ohmic conduction mechanism in high-dielectric-constant zro2 thin films,” *Journal of The Electrochemical Society*, vol. 152, 2005.

[22] T. Chiang and J. Wager, “Electron conduction mechanisms in insulators,” *IEEE TRANSACTION ON ELECTRON DEVICES*, vol. 65, no. 1, pp. 223–230, 2018.

[23] A. Bravaix, C. Guerin, V. Huard, D. Roy, J. Roux, and E. Vincent, “Hot-carrier acceleration factors for low power management in dc-ac stressed 40nm nmos node at high temperature,” 05 2009, pp. 531 – 548.

[24] P. Heremans, G. Van den Bosch, R. Bellens, G. Groeseneken, and H. Maes, “Temperature dependence of the channel hot-carrier degradation of n-

channel mosfet's," *IEEE Transactions on Electron Devices*, vol. 37, no. 4, pp. 980–993, 1990.

[25] S. Tyaginov, *Physics-Based Modeling of Hot-Carrier Degradation.* Springer International, 2015.

[26] K. Takahashi, K. Komatsu, T. Sakamoto, K. Kimura, F. Matsuoka, Y. Ishii, K. Egashira, and M. Sakai, "Hot-carrier induced off-state leakage current increase of ldmos and approach to overcome the phenomenon," in *2018 IEEE 30th International Symposium on Power Semiconductor Devices and ICs (ISPSD)*, 2018, pp. 303–306.

[27] T. Speranza, Y. Wu, E. Fisch, J. Slinkman, J. Wong, and K. Beyer, "Manufacturing optimization of shallow trench isolation for advanced cmos logic technology," in *2001 IEEE/SEMI Advanced Semiconductor Manufacturing Conference (IEEE Cat. No.01CH37160)*, 2001, pp. 59–63.

[28] R. Parthier, *Messtechnik, Grundlagen und Anwendungen der elektrischen Messtechnik.* Springer Vieweg, 2016.

[29] J. Fellner, "A one time programming cell using more than two resistance levels of a polyfuse," in *Proceedings of the IEEE 2005 Custom Integrated Circuits Conference, 2005.*, 2005, pp. 263–266.

[30] M. Kimura, "Field and temperature acceleration model for time-dependent dielectric breakdown," *IEEE Transactions on Electron Devices*, vol. 46, no. 1, pp. 220–229, 1999.

[31] D. Doane and L. Seward, "Measuring skewness: A forgotten statistic?" *Journal of Statistics Education*, vol. 19, pp. 6–7, 2011.

[32] R. Chattamvelli, Rajan; Shanmugam, *Statistics for scientists and engineers.* John Wiley Sons, 2015.

[33] L. Giannuzzi and F. Stevie, *INTRODUCTION TO FOCUSED ION BEAMS Instrumentation, Theory, Techniques and Practice.* springer, 2005.

[34] A. Gehring and S. Selberherr, "Modeling of tunneling current and gate dielectric reliability for nonvolatile memory devices," *IEEE TRANSACTIONS ON DEVICE AND MATERIALS RELIABILITY*, vol. 4, no. 3, pp. 306–319, 2004.

[35] U. Hilleringmann, *Silizium Halbleitertechnologie Grundlagen mikroelektronischer Integrationstechnik.* springer, 2014.

[36] Y.-W. L., W.-H. C., Y.-C. C., C.-S. C., and C.-D. C., "Effect of focused ion beam deposition induced contamination on the transport properties of nano devices," *Nanotechnology*, vol. 26, no. 5, 2015.