Final Project
Report

Study of Bias Temperature Instability in 180 nm pMOS

Submitted in partial fulfillment of
the requirements for

Summer Internship

Submitted by
Dibyan Goswami
B.Tech-School of Electrical Engineering, VIT Vellore

Under the guidance of
Dr. Shammi Verma
Scientist/Engineer, QARG

Semiconductor Laboratory
DEPARTMENT OF SPACE, GOVERNMENT OF INDIA
Sector 72, S.A.S Nagar, Punjab, India – 160 071

June 2022

Declaration

This is to certify that this is a bonafide record of the Summer Internship undertaken by Dibyan Goswami-20BEE0059 at Semiconductor Laboratory, Mohali during June 2022 in partial fulfilment of the requirements of the degree of Bachelor of Technology in Electrical and Electronics Engineering from VIT Vellore.

Dibyan Goswami
(Student Intern)

Dr. Shammi Verma
(Project Mentor)

Date: 29 June 2022

Acknowledgments

Contents

List of Figures

Chapter 1

Introduction to Project

If software was a person then Silicon would be its water. Si has made the software revolution possible. Semiconductors allow us to realise boolean logics on the nanoscale and ultimately allow us to integrate several of them together to build an Integrated Circuit. ICs play an important role in control applications everywhere. There are ICs everywhere around us, from the air-conditioner cooling our room to the car taking us to office. It is therefore, of importance to us to understand semiconductors, their working and their shortcomings to become good hardware engineers who can develop high quality ICs that not just ease our lives but make scientific and technological innovations easier to achieve. In this report, we start by discussing about Silicon: its fabrication process. We then talk about CMOS fabrication and MOSFET operation. Towards the end we discuss the failure modes of a MOSFET and the short channel effects. Finally, we discuss about NBTI and NBTI testing. Negative Bias Temperature Instability is a failure mode in MOSFETs. NBTI manifests itself in the form of increase in threshold voltage and a consequent decrease of drain current and therefore transconductance of the MOSFET. NBTI is of a concern in pMOSFETs that operate under negative voltage and has been on the rise with incorporation of Nitrogen in the oxide layer and decrease in MOSFET size with every technological node. It is therefore of importance to test fabricated pMOSFETs for NBTI and understanding the test conditions and how to identify a faulty device after testing.

Chapter 2

Fabrication Process of CMOS

2.1 Silicon Processing

2.1.1 Why Do We Use Silicon

We prefer Si for fabricating MOSFETs. Si is an ideal bulk semiconductor as it is easily available in large quantities, it has a bandgap of 1.12 eV which is greater than that of Germanium (0.67 eV) which means it absorbs and emits radiation in the visible region. Si can be worked at higher temperatures while Ge crystals disintegrate upon heating. Si also forms a protective SiO2 layer simply on heating which is hydrophobic in nature and eases the process of fabricating MOSFETs by acting as a barrier between bulk Si and Gate contact.

Figure 2.1: Silicon is widely available in nature in the form of sand

2.1.2 Process of obtaining Silicon

First step is simple, we reduce silica sand obtained from nature with coke to obtain Metallurgical Grade Silicon (MGS) of 98We then treat MGS with hydrochloric acid to obtain Trichlorosilane. Impurities in MGS are converted to

FeCl3, BCl3, PCl3/ PCl5 which can be removed by distillation. Trichlorosilane on further hydrogenation produces Si. This gives us highly pure silicon called Electronic Grade Silicon (EGS).

2.1.3 Silicon Wafering

We use the EGS obtained to produce wafers. We melt EGS in a crucible and inject a Si seed of a known orientation and rotate it and pull out the silicon melt at a slow rate. It is necessary for it to be a slow rate so that the silicon melt crystallises into the crystal of the same orientation as the seed without deformities and defects.

Figure 2.2: A Silicon ingot

Once the silicon ingot is obtained, we confirm crystal orientation using X-ray diffraction. We then grind a flat or notch on the ingot for the desired crystal orientation and saw it into wafers. We remove surface taper by lapping. As Si is brittle it is essential for there to be no cracks as they propagate through the entire surface. Therefore, we smoothen the edges using edge smoothening. After this we use laser scribing to ID the wafer and then etch it to remove damages caused by lapping and then polish it. Polishing allows us to thin the wafers further and remove contaminants on the surface.

Figure 2.3: A Silicon wafer

2.2 CMOS Fabrication

In CMOS fabrication we start with a Si layer that is either p-type or n-type. Let us consider a p-type substrate in our case.

Figure 2.4: bare wafer

It is heated in a furnace to form a protective layer of SiO2.

Figure 2.5: Oxide coating

We then add a photoresist over this layer.
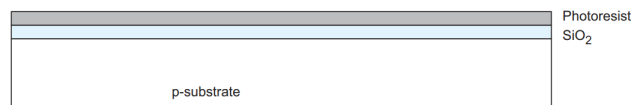
Figure 2.6: Photoresist coating

Say we wish to create an n-well in the bulk Si for fabricating the pMOS. We create the mask for n-well and expose the photoresist through it making it soluble in that masked area.

Figure 2.7: Masking and Etching

We etch out SiO2 layer using HF to expose the bulk Si.
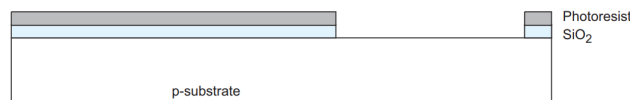
Figure 2.8: Oxide Etching

4

The remaining photoresist is removed using piranha etch.

Figure 2.9: PR washing

We then add the n-type dopants by diffusion or ion implantation. Ion implantation provides us with the benefits of anisotropy and no lateral diffusion of dopants and better control over dopant depth.
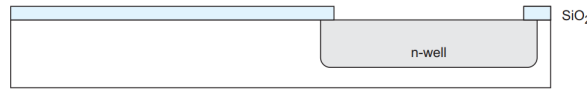
Figure 2.10: Ion Implantation

Excess SiO2 is removed using HF.

Figure 2.11: Etching

This gives us a Si wafer with an n-well and now we are left with generation of metal contacts and source and drain diffusion. We once again heat the Si wafer to form SiO2. Heavily doped polysilicon is coated over SiO2 by Chemical Vapour Deposition. The wafer is heated in the presence of SiH4 to obtain wafer coated with polysilicon.
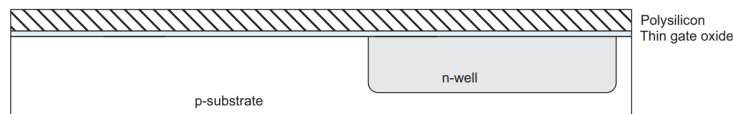
Figure 2.12: Polycrystalline silicon growth

We pattern the area where gate contacts are to be present by using photoresist and a polysilicon mask and etch out the polysilicon and SiO2 layer from elsewhere. Leaving the gate contacts atop the oxide layer.
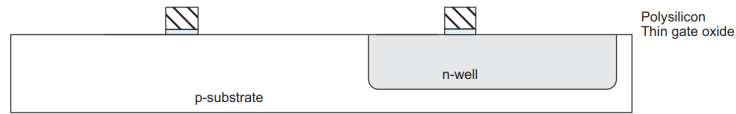
Figure 2.13: Gate Formation

We reheat the wafer to obtain a layer of SiO2 again and add masks in the regions where we require heavily doped n-regions for our nMOS and add dopants using the same previous method of exposure to light and etching.
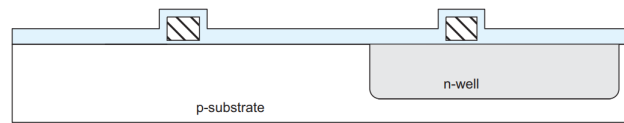
Figure 2.14: Oxide Growth

We add masks to expose the region where n-type dopants are needed for the nMOS and etch out the oxide.
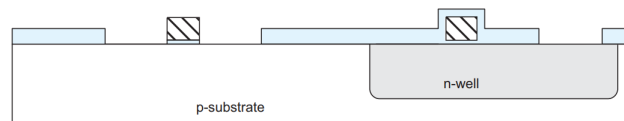
Figure 2.15: Source-Drain Masking and Etching

We add the n-type dopants by ion implantation. Traditionally, the dopants were diffused and therefore these regions are referred to as n-diffusion regions.
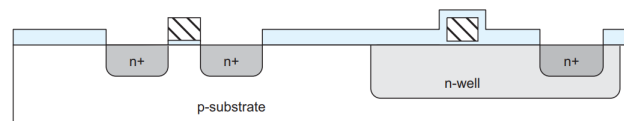
Figure 2.16: Dopant Implantation

We then strip off the protective oxide layer.
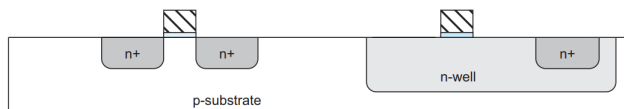
Figure 2.17: Oxide Etching

6

Similarly, we heavily dope the n-well with p type dopants to form the pMOS.

Figure 2.18: Fabricated CMOS

We then add masks for the metal to expose the region where metal contacts are needed.
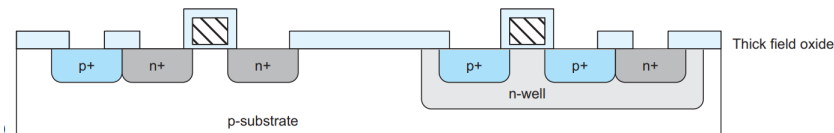
Figure 2.19: Fabricated CMOS

We sputter the metal over the entire wafer by blasting aluminium into vapour that evenly coats over the wafer. We use an alloy of Al and Cu. We use Cu so that Al doesn't undergo tapering due to current flow which is a common failure mode of MOSFET called Electromigration. We remove the remaining metal from all areas by plasma etching leaving metal only where the contacts are required.
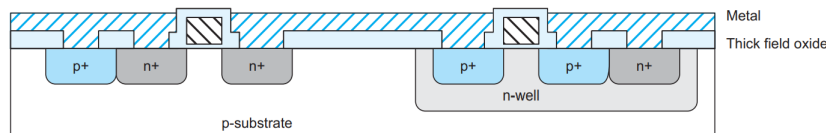
Figure 2.20: Fabricated CMOS with metal contacts

Chapter 3

MOSFET Operation

MOSFET stands for Metal Oxide Semiconductor Field Effect Transistor.
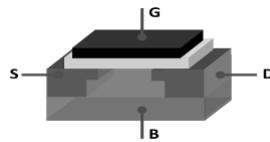
Figure 3.1: Simple diagram of a MOSFET

A MOSFET comprises of an Oxide layer sandwiched between a metal contact (or Polysilicon) and a semiconductor (Bulk Si). It controls current flow using an electric field applied via the Gate terminal and is therefore called a Field Effect Transistor. Carriers flow from the Source to Drain which are heavily doped semiconductors diffused into the bulk.

3.1 Working of a MOSFET and IV Characteristics

Let us consider a MOSFET with p-type bulk Si and heavily doper n-type Si at the source and drain. Initially, let gate and substrate voltage be 0 and VDD be applied to drain. As source-substrate and drain-substrate junctions are reverse biased, no current flows. We refer to this region as sub-threshold region

N-Channel E-MOSFET Structure
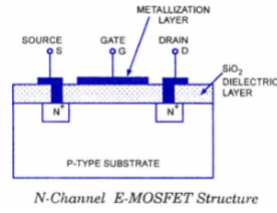
Figure 3.2: An N-channel enhancement mode MOSFET

Upon application of a gate voltage, the positive voltage attracts electrons and repels holes in the substrate. Repelled holes and incoming electrons recombine to form Si atoms. However, after a certain threshold is crossed, the accumulation of electrons overwhelms recombination. At and beyond this value of threshold gate voltage, electrons form a majority below the gate oxide and form a thin channel.
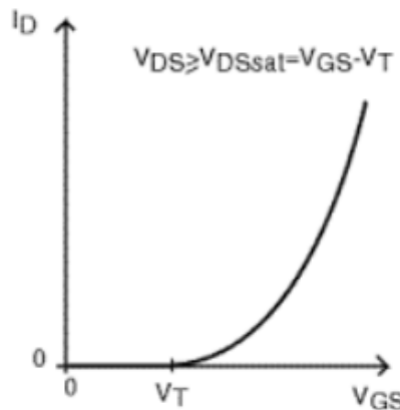
Figure 3.3: Drain current plotted as a function of gate voltage

As this channel is composed of electrons, it is known as an n-type channel and hence an n-type MOSFET. Also, the gate voltage enhances the concentration of n-type carriers and it is therefore an enhancement-mode MOSFET. Beyond the threshold voltage, there is a very low-resistance path for the drain current that is produced by the voltage at drain. The channel acts as a linear resistor and therefore drain current increases linearly with drain voltage. This is known as non-saturation region.
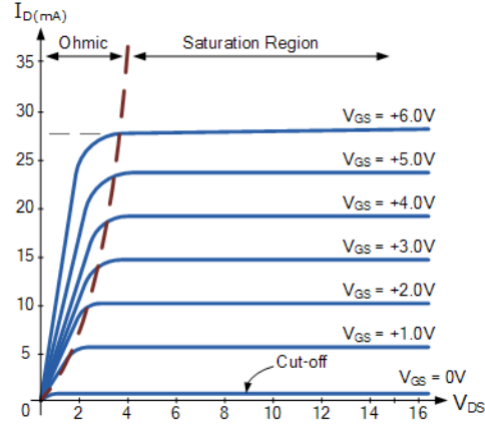
Figure 3.4: Drain current plotted against drain voltage

However, with an increase in drain voltage, the potential drop between the source and drain decreases. This reduction in potential drop across the channel lowers its width. Also, increase in drain voltage attracts n-type carriers away from the channel producing a pinch off and increasing the width of depletion region. A wider depletion region imposes greater resistance and thus drain current remains constant in spite of increasing voltage. This is known as saturation region.

3.2 CV Characteristics
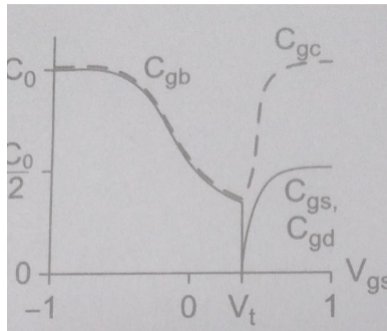
Figure 3.5: Capacitance plotted against gate voltage

Under no gate bias, the capacitance of the MOSFET equals the oxide capacitance formed between metal contact and the substrate. capacitance due to depletion regions is zero as none are formed. As gate bias is increased and

10

channel formation begins, the effective distance between the plates increases reducing the capacitance. Upon formation of channel, for low values of drain voltage, capacitance between gate to source and gate to drain is equal hence C0/2 as they are in series.

Chapter 4

Reliability and Quality

4.1 Reliability Physics

4.1.1 Introduction and General Thermodynamics

Quality refers to the property of the CMOS to work as intended without failing. Reliability is quality measured over time. Thus, reliability is a measure of the fabricated CMOS to function properly during its life.
Our fabricated devices are prone to several failure mechanisms, for example, capacitor leakage in an IC leads to its failure. Reliability physics studies the kinetics of failure mechanisms whereas reliability engineering refers to the practices to reduce failure probability such as: proper design rules, material selection, manufacturing guidelines etc.

Figure 4.1: A capacitor that has broken down due to dielectric leakage

As a consequence of Moore's law, there is a steady decrease in device

size with every technology node which implies a higher current density and electric field. Thus, even during normal use a device is close to its breakdown state as the operating voltage isn't scaled linearly with the size. Reliability studies of a device are therefore of great importance. One might wonder, why does a device fail? In fact, most of the devices that we use are metastable i.e they are stable now but are prone to fail in the future. Devices fail because according to thermodynamics, all systems tend to assume the most stable state (possessing zero potential energy) however input energy is required to bring about this change. The Gibb's potential is a good measure of a device's internal energy. Let us consider a capacitor, electric field does work against conservative forces to polarize the dielectric. This increases the Gibb's energy of the dielectric and it therefore wishes to assume a state of lower Gibb's energy which it achieves via dielectric breakdown that lowers its internal energy to zero thus allowing it to assume a stable state.

4.1.2 Hazard function

Hazard function, h(t) is the instantaneous failure rate of a device. It tells us the number of devices that will fail, mean time-to-failure and expected number of failures in an interval. There are different types of hazard functions: constant, linearly increasing, decreasing etc, representing differing failure rates.

For a given data, we have f(t) which is our p.d.f for failure that tells us number of failures between t and t+dt, F(t) as cumulative fraction that tells us total number of failures until t. R(t) is a measure of devices that work till time t. It is a measure of reliability. R(t)+H(t)=1 always. Hazard function is ratio of failure pdf to R(t).

4.2 Reliability Statistics

In order to determine the reliability of devices, statistics plays a crucial role as it allows us to determine a device's lifetime, plot and visualize device data and study about their reliability, failure etc. There are three curves of importance to us; Gaussian and Lognormal distribution: used for fitting reliability data for failure due to processes like corrosion, and Weibull distribution for fitting data for weakest link failure modes.

4.2.1 Gaussian, Lognormal and Weibull Distribution

Gaussian distribution is widely used for time-to-failure modelling of devices that fail due to failure modes such as corrosion. We determine the cum-fraction for each data point after arranging them in ascending order and plot Z (standard deviation) as a function of time. We get a straight-line graph with Z=0 corresponding to median (time at which 50
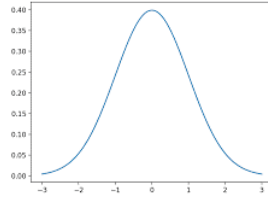
Figure 4.2: A gaussian distribution

Using the data, we can calculate instantaneous survival failure rate and average failure rate. Lognormal distribution is plotted similarly but by using a different plotting function.

Figure 4.3: A lognormal distribution

Weibull function is used for weakest link modelling and is characterized by two parameters: time-to-failure and shape parameter. Differing values for shape parameter give us different curves allowing us to determine device stability.
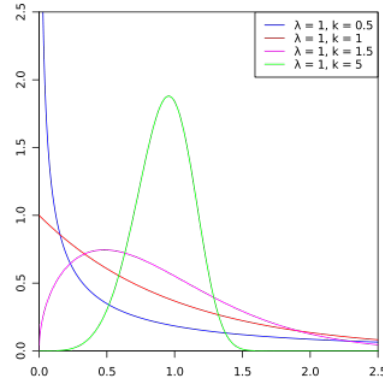
Figure 4.4: Weibull distribution. Differing values of shape parameter give us different graphs gaussian, exponential etc

4.3   Bathtub Curve for Reliability
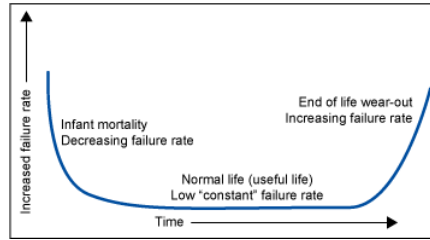
Figure 4.5: The bathtub curve

Bathtub curve is a curve that fits hazard function or failure rate as a function of time. It has three regions: EFR (Early-Failure Rate), IFR (Intrinsic-Failure Rate) and the Wear-out region. EFR: this region corresponds to an exponentially decreasing hazard function. This region symbolizes a very high-failure rate reducing drastically with time. In electronics, this may be due to manufacturing defects resulting in low-breakdown strength etc. A device can last for a year in the EFR under operating conditions. In order to remove the device from EFR region, we put it under increased stress to accelerate the time-to-failure process and this period is known as burn-in. The burn-in period is used to push the device through a year of its lifetime, if it is defective it fails during the burn-in period. The burn-in allows to remove devices from EFR and enter IFR.

IFR: this region corresponds to a low and constant hazard function. The failure rate is low and constant for a long period of time. Failure can occur

in this region due to small defects. Wear-out region: In this region, the failure rate increases monotonically with time. Failure occurs due to wear-out mechanisms resulting from long-term usage, for example, dielectric failure. To understand the bathtub curve, we can take the example of human life. When we are born, the first year as an infant is very important and utmost care is taken. This is because infants have a very high mortality rate. Infant Mortality Rate is a concern for all societies. We take utmost care of the infant when it is born to ensure that the baby is healthy as any form of disease can be fatal. Once the baby grows up, the mortality rate is relatively very low and we don't have to worry about our health as we did for the baby. There are few chances of us being struck by a fatal disease, we live for a long time working healthily. But as we get older, our body becomes weaker and once again we become susceptible to ailments of the old age that keep getting worse with time. We were in the EFR as a baby, in the IFR as a boy and a man and in the Wear-out region as an old man.

Chapter 5

Failure Modes in MOSFETs

5.1 Failure Mechanisms

5.1.1 Time Dependent Dielectric Breakdown

TDDB refers to the wear-out of dielectric with time. It refers to the formation of a conductive path between the gate and substrate because of which the gate field can't control the drain current. TDDB depends on the number of defects in the gate oxide due to wafer processing, therefore to reduce it our gate oxide must be ultra-pure. TDDB occurs due to prolonged usage of the device and is not an immediate effect. It occurs for a device in the intrinsic region of the bathtub curve
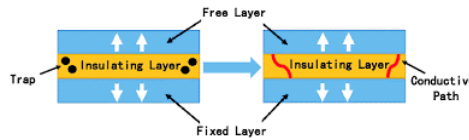
Figure 5.1: Traps contributing to electron tunnelling in TDDB

Sometimes, an electron can gain enough energy to overcome the Si-SiO2 interface potential barrier. Once this barrier is breached, electrons are accelerated through the dielectric breaking bonds thus creating traps. The traps increase the tunnelling current through the dielectric by providing a low barrier path across the dielectric to the electrons giving rise to a tunnelling current. This is a positive feedback process. The trap creation increases until the tunnelling current burns a hole in the dielectric.

5.1.2 Hot Carrier Damage

in a MOSFET, after pinch-off the lateral electrical field is restricted to the region between pinch-off and the drain contact. Thus, the electrical field increases as the length decreases increasing the drift current of the electron. This gives rise to hot electrons that may cause impact ionization, HCI, substrate leakage current and trap generation.

5.1.3 Electromigration

EM refers to the diffusion of metal ions along the direction of electron flow in a conductor. As electrons collide with an aluminium atom, the momentum transfer increases the probability of the atom moving parallel to the electron. As the aluminium atom vacates its initial position, it creates a vacancy and instead diffuses across the conductor to fill a vacancy due to a crystal defect. When the vacancies at the initial position of the aluminium atom are not replenished, it creates a defect there. The defects coalesce to produce a large void that eventually depletes the region of all aluminium atoms in the wire leading to circuit failure.

5.2 Short Channel Effects

Short-Channel Effects refer to the effects that arise in a MOSFET when the channel length is comparable to the size of the source and drain depletion region.

5.2.1 Drain Induced Barrier Lowering

DIBL refers to the lowering of threshold voltage due to high drain voltage. A high drain voltage reduces the barrier voltage at the source by increasing the forward bias effect allowing easy injection of electrons from the source into the channel. Thus, the drain voltage is inducing a lowering in barrier at the source. A higher drain voltage also reduces the charges the gate voltage must balance by assuming a greater share of it.

5.2.2 Velocity Saturation

in short channel devices, electric field is usually very high as voltage isn't scaled linearly with channel length. Drift velocity of an electron increases linearly with electric field. Beyond a certain value of electric field, the drift
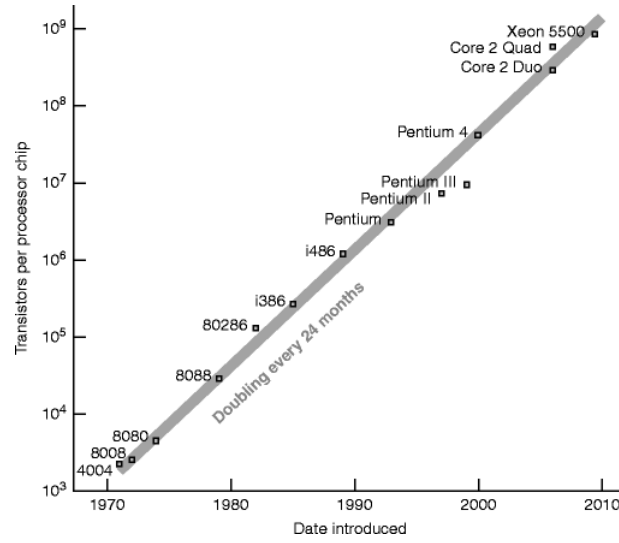
Figure 5.2: Graph depicting Moore's law. Transistors per chip v. date Short-channel effects are inevitable as a consequence of Moore's law

velocity saturates. An electron in a short-channel device is under high electric field stress. Thus, electrons attain their saturation drift velocity much before the drain saturation voltage. At saturation velocity, the drain current also saturates and therefore current saturation occurs. But current saturation occurs before drain saturation voltage indicating that current saturation occurs as a result of drift velocity saturation and not pinch-off. This reduces the reliability of the device as it no longer saturates at a known value of drain voltage.

5.2.3 Quantum Confinement

At very small device sizes, the device size and de-broglie wavelength of electron become comparable. This implies that laws of classical mechanics no longer apply to the electron and we must use quantum mechanics to determine their behaviour.

5.2.4 Hot Carrier Injection

a hot carrier is a highly energetic charge carrier. Hot carriers are generated by incidence of photons on the substrate that creates a hole and a highly energetic electron called a hot electron. In hot carrier injection, an electron is injected from the n-channel to the gate oxide dielectric increasing its threshold voltage by oxide charging resulting in poor gate control. Holes

travel out of the channel via the substrate giving rise to leakage current that can be measured to determine the extent of HCI. Hot electrons also strike the Si-H bonds at the interface generating traps that contribute to further dielectric degradation. HCI is more pronounced at low temperatures as collision probability is greatly reduced at higher temperatures due to increased vibration.

5.2.5 Impact Ionization

hot electrons can collide with other Si atoms generating electron-hole pair and increasing the drain current past saturation.

5.2.6 Mobility Degradation

as we scale down, vertical electric field increases causing electron scattering at the Si-SiO2 interface reducing mobility of carriers.

5.3 NBTI Mechanism

NBTI is an effect that shifts the threshold voltage for a pMOS when a negative gate bias voltage is applied at elevated temperatures. NBTI arises due to trapped charges at the interface and the oxide layer. The resulting threshold voltage rise leads to a decrease in drain saturation current. The threshold voltage shift increases with stress time and reduces as the stress is removed. This is because threshold voltage shift occurs due to interface and bulk trap charges, as the negative electric field is removed, the charges diffuse back to their initial positions restoring the broken bonds and reducing the threshold voltage shift. NBTI is more pronounced in (110) crystal lattices that have higher hole mobility due to greater density of Si-H bonds. Addition of Nitrogen induces additional hole traps in the bulk oxide. Thinner oxide layer too contributes to a greater electric field that ultimately leads to a more pronounced NBTI effect. Deuterium and Fluorine help reduce NBTI by replace the Si-H bond with a much stronger bond resulting in lesser trap generation. NBTI can be modelled by the reaction-diffusion and two-stage model. RD model describes the breaking of Si-H interface bonds by holes as a two-way process. Channel holes break the bonds that releases hydrogen that can diffuse through the oxide layer generating a hole trap. The diffusing hydrogen can also get trapped in process-induced bulk traps in the oxide layer. Electron and hole assisted tunnelling can further activate pre-existing traps.

Activation of these traps ultimately rises the threshold voltage, degrading the oxide and disrupting operation.

Chapter 6

NBTI testing

6.1 Pre-test

As NBTI is a prominent failure mode in 180 nm process, it is necessary for us to test our devices for failure under NBTI stress before sending them for further application. The first step in NBTI testing of a pMOS is determination of its threshold voltage. As NBTI leads to an increase in threshold voltage, we must know its nominal threshold voltage. We determine the threshold voltage of the pMOS by low-voltage of high-voltage testing.

After determining threshold voltage, we test the device under NBTI test. For this, we use a machine integrated with a software to ease the testing process. We insert all the packages on a DUT board that has pin-outs that are connected to external circuitry in the machine. A DUT board can host several devices allowing for statistical data to be acquired.

Before NBTI, we check if gate current and drain saturation current is within an acceptable limit by applying operating gate voltage to both gate and source. We also measure the off current of the MOSFET by applying zero gate voltage.

6.2 Test Conditions

We put the device under NBTI stress by applying a temperature of 150 C and a gate voltage around 10 percent above the operating gate voltage for a period of 168 hours. The drain and source voltage must be 0. This is know as symmetrical configuration. If the device shows a deviation of 10 percent from it's initially measured threshold voltage, it is considered failed.

While testing, we have to wait for a while before we actually make measurements after applying the stress and this period is referred to as stress soak

time. This is necessary to ensure uniform application of stress conditions on the device.

6.3 Test Results

Device ID	threshold voltage	leakage current

Chapter 7

Conclusion

The lot of pMOSFET was tested successfully under NBTI test and all the devices passed the test. The variation in threshold voltage was lesser than 10 percent indicatinmg that our devices are suitable for further use. The experiment yielded acceptable and expected values of threshold voltage and leakage current.

References

[1] Introduction to Microfabrication, Second Edition, `https://onlinelibrary.wiley.com/doi/book/10.1002/9781119990413`

[2] Physics of Semiconductor Devices `https://onlinelibrary.wiley.com/doi/book/10.1002/0470068329`

[3] CMOS VLSI Design: A Circuits and Systems Perspective `https://dl.acm.org/doi/10.5555/1841628`

[4] Short-Channel Effects `https://en.wikipedia.org/wiki/Short_Channel_Effects`

[5] Reliability Physics and Engineering `https://link.springer.com/book/10.1007/978-3-319-93683-3`

[6] Reliability Wearout Mechanisms in Advanced CMOS Technologies `https://ieeexplore.ieee.org/book/5361029`