

DISSERTATION

Advanced Characterization of the Bias Temperature Instability

ausgeführt zum Zwecke der Erlangung des akademischen Grades
eines Doktors der technischen Wissenschaften

eingereicht an der Technischen Universität Wien
Fakultät für Elektrotechnik und Informationstechnik
von

PHILIPP HEHENBERGER



Matr. Nr. 0025027

geboren am 12. Oktober 1980 in Wien, Österreich

Wien, im November 2011



Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

*It is unwise
to be too sure
of one's own wisdom.*

*It is healthy
to be reminded that
the strongest might weaken
and the wisest might err.*

Mahatma Gandhi



Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.



Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

To Veronika



Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Danksagung

Auch wenn diese Arbeit meinen Namen trägt, möchte ich anmerken, dass es ohne die Unterstützung vieler großartiger und hilfsbereiter Menschen nur schwer gewesen wäre, sie zu verfassen.

Zuallererst gebührt mein aufrichtiger Dank meinem Betreuer *Prof. Tibor Grasser*. Ich habe in meinem Leben nicht viele Menschen getroffen, die Wissenschaft mit solcher Begeisterung verfolgen und die ihre Leidenschaft in jeder Sekunde versuchen weiterzugeben. Sein unermüdlicher Eifer beim Lösen von Problemstellungen hat mich oft beeindruckt. Trotz immer vollen Terminkalenders war er stets erreichbar, ob persönlich, per Telefon oder per Mail. Im nachhinein am meisten dankbar bin ich ihm wahrscheinlich dafür, dass er mir vor vier Jahren die Universitätsassistentenstelle ans Herz gelegt hat. Damals konnte ich noch nicht erahnen, dass er mir damit die Tür zur Welt der Programmierung geöffnet hat, ohne die hier am Institut für Mikroelektronik vieles nicht möglich ist.

Prof. Martin Gröschl vom Institut für allgemeine Physik bin ich sehr dankbar, dass er sich wie schon bei meiner Diplomarbeit als Gutachter zur Verfügung stellt. Ich selbst bin im Herzen immer Physiker geblieben.

Weiters danke ich *Prof. Erasmus Langer* und *Prof. Siegfried Selberherr*, die in all den Jahren stets ein offenes Ohr für meine Anliegen hatten und die dafür gesorgt haben, dass es mir an meinem Arbeitsplatz an nichts fehlt, und das ist keine Selbstverständlichkeit. Ebenfalls möchte ich mich für das von ihnen geschenkte Vertrauen bedanken, dass sie mir als Assistent entgegengebracht haben. Auch *Manfred Katterbauer*, *Ewald Haslinger* und *Renate Winkler* bin ich für die vielen Kleinigkeiten, die man viel zu schnell vergisst, dankbar.

Ich muss schon jetzt zugeben, dass ich viele meiner Kollegen schmerzlich vermissen werde. Zu groß war ihr Einfluss, ihre Motivation, ihre Hilfen und vor allem die abwechslungsreichen Gespräche über Gott und die Welt. Natürlich möchte ich mich bei allen Kollegen bedanken, stellvertretend werde ich mein näheres Umfeld erwähnen.

Nachdem wir beinahe zeitgleich am Institut für Mikroelektronik angefangen haben, macht *Paul-Jürgen Wagner* hier den Anfang. Durch seine ausgezeichneten Englischkenntnisse hat diese Arbeit den letzten Feinschliff erhalten. Seine Fertigkeiten in Latex nicht hervorzuheben wäre falsch und für die Geduld, die er mir immer wieder entgegen gebracht hat, kann ich mich nicht genug bedanken.

Franz Schanovsky ist nicht nur ein begnadeter Programmierer, sondern wie Paul auch mein Zimmerkollege. Es ließ sich nicht vermeiden, dass ich Franz immer dann, wenn es Probleme gab, zu Rate gezogen habe. Was meist mit einem „Was gibt es, Philipp?“ anfang, war bereits nach kürzester Zeit mit einem „Freut mich, dass ich dir helfen konnte!“ aus der Welt geschafft. Manchmal schien mir, dass die bloße Anwesenheit von Paul oder Franz mein Problem gelöst hat. Wir drei haben uns in den letzten Jahren sehr zusammengelebt, und der subtile Humor, mit dem wir uns stets begegnet sind, hat mein Leben um viele Facetten bereichert.

Oskar Baumgartner und *Zlatan Stanojevic* danke ich für die vielen Hilfestellungen bei vsp, aber auch für so manche Kaffeepause. Mit *Wolfgang Gös* konnte ich mich jederzeit über charge-trapping unterhalten und ohne *Oliver Triebel* wären meine ersten Minimos-Gehversuche wohl erheblich schleppender verlaufen. Vielen Dank auch an *Markus Karner* und sein gesamtes gts-Team. Sie haben mich, wann immer nötig, unterstützt. Nicht zuletzt ist *Johann Cervenka* einer der wenigen, der um jede Tageszeit – ob am Institut anwesend oder nicht – innerhalb kürzester Zeit auf unser Netzwerk zugreifen kann, auch dafür danke.

Zwar nicht direkt mit unserem Institut verbunden, aber ein sehr gewissenhafter Wissenschaftler und gesegneter Tüftler ist *Hans Reisinger*. Ich durfte ihn zahlreiche Male in München besuchen, um dort an Bauteilen spezielle Messungen mit seinem eigens erbauten Equipment durchzuführen.

Eine weitere außergewöhnliche Erfahrung habe ich während einer Kooperation mit der Fudan Universität in Shanghai gemacht. Die Gruppe um *Prof. Ming-Fu Li* mit *ZhiYing Liu* und *WenJun Liu* hat mir Einblick in ihre Messmethoden ermöglicht und die Gastfreundschaft, die ich in China von wildfremden Leuten erfahren habe, hat meine Vorstellungen bei weitem übertroffen.

Bei unseren Kooperationspartnern möchte ich mich ebenfalls für die tatkräftige Unterstützung und das Bereitstellen von Messdaten bedanken. Stellvertretend seien hier *Thomas Aichinger* vom KAI und *Michael Nelhiebel* von Infineon Villach genannt. Von Belgien aus hatten *Ben Kaczer* und *Jacopo Franco* von imec wesentlichen Anteil am Vorankommen meiner Forschung.

Schließlich möchte ich mich bei meinen Freunden bedanken, die immer an mich geglaubt haben, selbst dann, wenn ich manchmal am Gelingen der Arbeit gezweifelt habe. *Florian Weinwurm* und *Matthias Dilger* hatten jederzeit einen Kaffee und ein „Relax!“ für mich übrig und als Leidensgenosse ist *Gernot Eller* in meinem privaten Umfeld quasi der einzige, der die Probleme eines Dissertanten kennt.

Auch ein Dank gebührt meinen Eltern und meiner Schwester, die mich immer unterstützt haben, ob finanziell oder mit Rat und Tat.

Zuletzt gebührt mein wichtigster Dank meiner Frau, die hinter mir steht. Ihre unendliche Geduld, die Kraft mich zu motivieren und ihre Art Dinge gelassen zu sehen, beeindruckten mich jeden Tag aufs Neue - Danke Veronika.

Sollte ich jemanden vergessen haben hier namentlich aufzuführen, so tut es mir leid. Auch all jenen ein herzliches Danke.

Kurzfassung

Die steigende Nachfrage nach elektronischen Produkten erfordert eine ständige Optimierung der Massenproduktion. Die Halbleiterindustrie als größter Zulieferer der Elektroniksparte erreicht dieses Ziel, indem sie die mikroelektronischen Bauteile immer weiter verkleinert. Als Folge dieser Verkleinerung wird die Zuverlässigkeit von elektronischen Komponenten wie dem Metall-Oxid-Halbleiter Feldeffekttransistor (*MOSFET*¹) ein immer ernstzunehmenderes Thema. Die *bias temperature instability* (BTI) stellt eine solche Herausforderung dar. Sie tritt auf, wenn das *gate* eines Transistors unter erhöhter elektrischer Spannung steht. Die darauffolgende Veränderung von diversen Transistorparametern, wie zum Beispiel der Einsatzspannung, wird weiters durch eine erhöhte Temperatur verstärkt. Eine elementare Herausforderung, die zum Verständnis von BTI beitragen soll, besteht darin, dass sich die durch Stress degradierten Parameter nach Beendigung desselben wieder erholen. Man nennt diesen Vorgang auch Relaxation.

In dieser Arbeit werden die Auswirkungen von sowohl negativen als auch positiven Gatespannungen ausführlich mit unterschiedlichen Messtechniken erfasst. Neben den in der Halbleiterindustrie gebräuchlichen handelsüblichen Messinstrumenten werden mitunter auch komplett in Eigenregie entwickelte und gebaute Messinstrumente verwendet. Leider gibt es kein perfektes Setup für die Charakterisierung von BTI und jedes einzelne Equipment hat spezifische Vor- und Nachteile. Basierend auf bereits existierenden Modellen zur Beschreibung von BTI wird gezeigt, dass die Zeitverzögerung, die bei der Messung erfolgt, einen großen Einfluss auf die Beschreibung der Degradation und somit auf die vom Hersteller geschätzte Lebensdauer von Bauteilen hat. Auch aus diesem Grund wird die Relaxation hier genau untersucht. Nachdem unterschiedliches Equipment zur Charakterisierung von BTI verwendet wird und es bisher leider keine allgemein gültigen Spezifikationsrichtlinien gibt, erschwert das den Vergleich von Messdaten. Weiters unterscheidet sich die Nachbehandlung der Messdaten von Messroutinen, was mitunter eine sehr heikle Angelegenheit darstellt.

Der einfachste Weg, die Anfälligkeit eines Bauteils für BTI zu bestimmen, besteht darin, zuerst eine Referenzmessung der betroffenen Messgröße vorzunehmen. Nach erfolgtem Stress wird dann die Veränderung der Messgröße evaluiert. Dieser Vorgang wird Messen-Stressen-Messen (MSM) genannt. Durch wiederholtes abwechselndes Messen und Stressen können beliebig viele solcher Relaxationssequenzen aufgenommen werden. Die sehr kurze Zeitverzögerung der Messroutine, die Unempfindlichkeit gegenüber der Beweglichkeitsänderung in Kanal eines MOSFET's und die Möglichkeit eine ungestresste Referenz zu erhalten, spricht für diese Technik im Vergleich zu anderen, obwohl die Stressphase bei MSM nicht aufgezeichnet werden kann. Um letzteres Handikap zu beseitigen, kann MSM mit der *on-the-fly* Methode, die eine Aufzeichnung der Stressesequenz ermöglicht, kombiniert werden.

¹Kursive Wörter stellen englisches Vokabular dar, dessen Übersetzung nach Meinung des Authors keinen Sinn macht.

Ein weiteres Hauptaugenmerk der Arbeit liegt darin, die Bedeutung von Kurzzeit- und Langzeitverhalten der Relaxation zu untersuchen. Obwohl BTI schon seit Jahrzehnten bekannt ist, kam die Erkenntnis, dass Relaxation über einen logarithmisch gesehen großen Zeitraum stattfindet, erst vor wenigen Jahren. Das ist auch der Grund, warum lange angenommen wurde, dass NBTI ausreichend genau durch Wasserstoffdiffusion in das Oxid erklärt werden kann. Durch die modellbedingte Rückdiffusion kann die Relaxation allerdings nicht zufriedenstellend erklärt werden. Weiters ist es unmöglich mit dieser Theorie das experimentell beobachtete Verhalten von BTI abhängig von Temperatur, elektrischer Feldstärke im Oxid und Frequenz zu beschreiben.

Neuere Modellansätze bedienen sich schneller Locheinfangprozesse und der langsameren Generation von Grenzflächenzuständen, um BTI zu erklären. Eine Vielzahl an Versuchen war notwendig, bis ein passender Mechanismus gefunden wurde, der in der Lage ist, das zeitlich sehr weite Relaxationsverhalten zu erklären. Immerhin muss das Modell über 12 Dekaden – und auch darüber hinaus – Gültigkeit besitzen. Ein passender Kandidat dafür ist die strahlungslose Multiphonon Theorie, mit der bereits $1/f$ -Rauschmessungen modelliert wurden. Diese Theorie basiert auf der Annahme, dass die Energie jedes Defektsystems durch ein adiabatisches Potential beschrieben werden kann. Durch das Anlegen von Stress kann ein thermodynamisch stabiles Defektpotential (1) gegenüber einem anderen höheren und deshalb unbesetzten Defektpotential (2) energetisch soweit angehoben werden, dass der thermodynamische Grundzustand von (1) nun über dem von (2) liegt. Dies ermöglicht einen Übergang von (1) zu (2), genannt Locheinfang. Während der Relaxation stellt sich wieder die energetisch niedrigere Defektkonfiguration von (1) ein, was einen Übergang zurück zu (1), also Lochabgabe, ermöglicht. Mittels eines solchen Modells konnte man bereits den Stufenprozess während der Relaxation von kleinflächigen MOSFETs als zeitlich diskret stattfindende Lochabgaben erklären. Weiters kann vorausgesetzt werden, dass größere MOSFETs auch eine größere Anzahl an Defekten aufweisen. Für diese Defekte kann man ferner unterschiedliche Eigenschaften, wie die energetische oder lokale Position im Oxid annehmen. Damit ist es möglich, Messungen bei unterschiedlichen Temperaturen, elektrischen Feldstärken im Oxid und auch Stresszeiten mittels der strahlungslosen Multiphonon Theorie zu modellieren, was ihre Gültigkeit in Bezug auf BTI unterstreicht.

Abstract

To keep up with the growing demand for electronic products, a continuous optimization of their mass production is necessary. The semiconductor industry as the main supplier in this market handles this optimization process via miniaturization of microelectronic devices, such as metal-oxide-semiconductor field effect transistors (MOSFETs) which are investigated here. As a consequence of the device shrinkage, reliability issues like the bias temperature instability (BTI) have become a serious topic. BTI happens when the gate is biased while the transistor is exposed to elevated temperatures. This process severely changes some of the transistor parameters, e.g. the threshold voltage. A fundamental challenge in understanding BTI is that the degradation is found to recover when the bias is removed.

In this thesis the characterization of both negative and positive BTI is studied by using different measurement techniques. In addition to commercial measurement tools also equipment conceived and built by Hans Reisinger from Infineon Technologies AG is used. Unfortunately, there is no perfect measurement technique and each one exhibits certain limitations. Based on existing modeling attempts it will be shown that the delay time of the measurement has a huge impact on the characterization of the degradation and therewith on the projected time to failure. This is also the reason why BTI recovery is investigated thoroughly here. Furthermore, BTI is generally not specified in a consistent way because of the different characterization equipments used. A comparison of different measurement routines will show that the postprocessing of measurement output data is a very delicate task.

The most simple way to determine the BTI sensitivity of a device is to first take a reference of the quantity that should be characterized, then stress the device for a well-defined time and afterwards measure the change of the quantity. This is called the measurement-stress-measurement (MSM) method. An extended version thereof alternately stresses the device and then monitors the degradation during recovery with ever increasing stress times. The advantages of the MSM technique compared to others are its very short measurement delay time, its insensitivity to mobility changes and the possibility to obtain an unstressed reference prior to stress. In an extended MSM setup the MSM technique is further combined with the on-the-fly method, which monitors the stress. This allows the observation of both stress and recovery.

A main task of this work is to study both short- and long-term stress and recovery behavior. Though BTI has been known for some decades, the finding that its recovery is spread over many time scales is quite new. This is also the reason why it was thought for a long time that NBTI can be sufficiently described via the diffusion of hydrogen generated at the interface. However, the well-known reaction-diffusion theory is not able to explain the recovery by back diffusion of hydrogen. Furthermore, the temperature, oxide electric field, and frequency dependencies during stress, which are all observed in experiments, can not be modeled by this theory.

Newer modeling approaches are based on faster hole trapping processes and slower interface state generation. It took many attempts to find a possible mechanism that is able to explain the wide time range of the recovery which sometimes exceeds even 12 decades in time. Such wide distributions of time constants have already been observed during the analysis of $1/f$ -noise spectra. Consequently, the models previously used for the explanation of $1/f$ -noise were taken as a starting point. In its extended form the defects are described by adiabatic potentials, which eventually determine the non-radiative multi-phonon (NMP) transitions between the various defect states. Upon the application of BTI stress the initial defect potential is shifted in energy and a transition into another defect configuration is favored. During recovery the transition back into its initial configuration is favored in turn. By such a mechanism it was already possible to explain the step-like recovery behavior of small-area devices by hole emission of single defects.

Large-area devices can also be modeled by using the same NMP theory with the only difference that more defects are necessary to describe BTI. These defects are assumed to exhibit different energies and distances inside the oxide. The correct description of measurement data that includes different temperatures, oxide electric fields, and stress times finally supports the validity of the NMP model for BTI.

Contents

Danksagung	i
Kurzfassung	iii
Abstract	v
List of Abbreviations	xi
List of Symbols	xiii
1 Introduction	1
1.1 Historical Background	1
1.2 BTI – Causes and Impacts	2
1.3 Modeling BTI with Defects	3
2 Measurement Methods	5
2.1 Measurement-Stress-Measurement	6
2.1.1 Monitoring I_D at V_{TH}	6
2.1.2 Direct Monitoring of V_{TH}	7
2.1.3 Extended-Measurement-Stress-Measurement Setup	7
2.2 Transfer-Characteristics	9
2.2.1 Fast Pulsed $I_D(V_G)$ -characteristics	9
2.2.2 Improved Method of Reisinger	9
2.3 On-The-Fly (OTF)	11
2.4 Charge Pumping	14
2.5 On-the-Fly Fast Charge Pumping	14
2.6 Capacitance Voltage Profiling	15
3 Previous Modeling Attempts	19
3.1 Reaction Diffusion Model	19
3.1.1 Stress Phase	19
3.1.2 Back Diffusion of Hydrogen during Recovery	21
3.2 Extensions of the Reaction-Diffusion Model	21
3.2.1 Dispersive-Reaction-Rate Models	22

4	Two Components Contributing to Bias Temperature Instability	25
4.1	Universality of BTI recovery	26
4.2	Assumption of a Permanent Component	27
4.2.1	Temperature and Voltage Dependence of Universal Law	29
4.2.2	Measurement Delay	29
4.3	ΔV_{TH} versus ΔV_{θ}	31
4.4	Conclusion	33
5	Pulsed BTI Measurements	35
5.1	Pulsed $I_D(V_G)$ -Characteristics	37
5.2	Further Data Extraction Options	38
5.2.1	Determination of the Fitting Region	39
5.2.2	Impact of the Pulse Amplitude	39
5.2.3	Varying Pulse Rise/Fall Times	40
5.2.4	Consequences	41
5.3	Experimental Identification of Defects	42
5.4	OFIT versus CP	42
5.5	Analysis of the OFIT Technique	43
5.5.1	Dependence on Gate Voltage Low-Level	44
5.5.2	Hysteresis due to Stress	44
5.6	Extrapolation of Oxide Trap Contribution	45
5.7	Simulation of the Charge Pumping Current	46
5.8	Results	48
5.9	Conclusion	49
6	Short-Term NBTI	51
6.1	Gate Pulse Settings	52
6.2	Data Extraction	53
6.2.1	Offset	54
6.2.2	Initial Measurement as Reference	54
6.2.3	Gate Voltage Criteria	55
6.2.4	Brute-Force Truncation of the Transient	56
6.2.5	Final Setting of Parameters	56
6.3	Logarithmic Stress Behavior	57
6.3.1	Used Samples and Stress Conditions	57
6.3.2	Temperature Scaling	57
6.3.3	Voltage Scaling	58
6.3.4	Oxide Thickness Scaling	58
6.3.5	Extracted Prefactors	58
6.4	Power-Law Stress Behavior	59
6.5	Relaxation Behavior	60
6.6	Fast Ramp versus Fast- V_{TH} -Method	62
6.7	Conclusions	62

7	Relaxation of Negative/Positive BTI	63
7.1	Raw Measurement Results	65
7.2	Schematic Recovery Behavior	66
7.3	Extraction Routine	66
7.4	Discussion of the Experimental Output	66
7.4.1	Stress Time Component	67
7.4.2	Oxide Electric Field Component	68
7.5	Short-Term and Long-Term Relaxation	68
7.5.1	Entire Relaxation	69
7.5.2	Change in ΔV_{TH}	70
7.6	Emission Time Constants	71
7.7	Conclusions	75
8	Latest Modeling Attempts - Hole Trapping	77
8.1	Rate Equations	78
8.2	Elastic Hole Trapping	82
8.3	Coupled Double-Well Model	82
8.4	Two-Stage Model	83
8.5	Multi-Phonon Emission	85
8.5.1	Approximation of the Vibronic Transition	86
8.5.2	Radiative Multi-Phonon Emission	86
8.5.3	Non-Radiative Multi-Phonon Theory	88
8.6	Conclusion	89
9	Modeling NBTI in High-k SiGe pMOSFETs	91
9.1	Inverse Modeling	92
9.2	Multi-State Defect Model	93
9.2.1	Distribution of Defects	96
9.2.2	Reservoir of Holes - Classical vs. Quantum Mechanical Description	98
9.3	Results	99
9.4	Conclusions	101
10	Summary and Outlook	103
A	Extracting V_{θ} Based on the Level 1 model	107
A.1	OTF1	107
A.2	OTF2	108
A.3	OTF3	108
B	Ideal MOS Capacitor	109
B.1	Surface Space Charge Region of an n-Type MOS Capacitor	109
B.2	Results for p-Type Semiconductors	113
C	Diffusion-Limited Stress Phase of the Reaction-Diffusion Theory	115

D Multi-Phonon Emission	117
D.1 Radiative Multi-Phonon Emission	117
D.2 Non-Radiative Multi-Phonon Process	118
Bibliography	121
Own Publications	135
Curriculum Vitae	137

List of Abbreviations

BTI	bias temperature instability
CMOS	complementary metal-oxide-semiconductor
CP	charge pumping
$C(V)$	capacitance as a function of voltage
DC	duty cycle or direct current
DFT	density function theory
DSO	digital storage oscilloscope
DUT	device under test
eMSM	extended measurement-stress-measurement
EOT	effective oxide thickness
FPM	fast pulsed measurement
FC	Franck-Condon
FD	Fermi-Dirac
HCI	hot carrier injection
$I_D(V_G)$	transfer characteristic
LSF	line-shape function
MB	Maxwell-Boltzmann
MOS	metal-oxide-semiconductor
MOSFET	metal-oxide-semiconductor field effect transistor
MPE	(radiative) multi-phonon emission
MSM	measurement-stress-measurement
NBTI	negative bias temperature instability
NMP	non-radiative multi-phonon
OTF	on-the-fly
OFIT	on-the-fly (charge pumping) interface traps
PBTI	positive bias temperature instability
RD	reaction-diffusion
RTN	random telegraph noise
SPICE	Simulation Program with Integrated Circuit Emphasis
SRH	Shockley-Read-Hall

List of Symbols

Symbol	Unit	Description
C_{ox}	F m^{-2}	Areal gate oxide capacitance
E_{A}	eV	Activation energy
E_{B}	eV	Binding energy
ΔE_{B}	eV	Thermally activated barrier
E_{c}	eV	Conduction bandedge energy
E_{f}	eV	Fermi energy
E_{ox}	V m^{-1}	Oxide electric field
E_{s}	V m^{-1}	Electric field at the surface
E_{v}	eV	Valence bandedge energy
E_i	eV	Energy of defect state i
E_{ij}	eV	Energy difference $E_i - E_j$
$\epsilon_{i'}$	eV	Energy barrier for the metastable transition $i \rightarrow i'$
ϵ_{ij}	eV	Energy barrier for the transition $i \rightarrow j$
ϵ_{r}	1	Relative permittivity
g_{m}	A V^{-1}	Transconductance
I_{cp}	A	Charge pumping current
I_{D}	A	Drain current
$I_{\text{D},0}$	A	Initial drain current
ΔI_{D}	A	Drain current shift
$I_{\text{D},\text{lin}}$	A	Linear drain current
I_{TH}	A	Drain current criterion
k_{f}	s^{-1}	Forward transition rate
k_{r}	s^{-1}	Reverse transition rate
k_{ij}	s^{-1}	Transition rate for $i \rightarrow j$
L	m	Gate length
μ_{eff}	$\text{m}^2 \text{V}^{-1} \text{s}^{-1}$	Effective mobility
N_{it}	m^{-2}	Number of interface states per area
N_{ot}	m^{-2}	Number of oxide traps per area
N_{v}	m^{-3}	Effective valence band weight
Q_{it}	C m^{-2}	Interface charge per area
Q_{ot}	C m^{-2}	Oxide charge per area
Q_{s}	C m^{-2}	Surface charge density
q_i	m	Reaction coordinate equilibrium of state i

Symbol	Unit	Description
$S\hbar\omega$	eV	Relaxation energy
σ_p	m ²	Cross section of holes
τ_c	s	Capture time constant
τ_e	s	Emission time constant
t_M	s	Measurement delay time
t_{ox}	m	Oxide thickness
t_P	s	Pulse period
t_{rel}	s	Relaxation time
t_{str}	s	Stress time
V_D	V	Drain voltage
V_G	V	Gate voltage
V_{rel}	V	Gate relaxation voltage
V_{str}	V	Gate stress voltage
V_i	V	Adiabatic potential of state i
V_{TH}	V	Threshold voltage
$V_{TH,0}$	V	Initial threshold voltage
ΔV_{TH}	V	Threshold voltage shift
$v_{th,p}$	m s ⁻¹	Thermal velocity of holes
V_θ	V	Threshold voltage after the compact model
ΔV_θ	V	Threshold voltage shift after the compact model
$\Delta V_\theta^{OTF,x}$	V	Threshold voltage shift by means of OTF,x
W	m	Gate width
ω_i	s ⁻¹	Vibronic frequency of state i
x_T	m	Oxide trap depth (referenced to interface)

Physical Constants

$\epsilon_0 = 8.854\,187\,818 \cdot 10^{-12} \text{ F m}^{-1}$...	Vacuum permittivity
$h = 1.054\,571\,726 \cdot 10^{-34} \text{ J s}$...	Planck constant
$\hbar = h/2\pi$...	Reduced Planck constant
$k_B = 1.380\,648\,813 \cdot 10^{-23} \text{ J K}^{-1}$...	Boltzmann constant
$q_0 = 1.602\,176\,565 \cdot 10^{-19} \text{ C}$...	Elementary charge

Chemical Symbols

H^0	...	Atomic hydrogen
H_2	...	Molecular hydrogen
Si^\bullet	...	Silicon dangling bond
$Si-H$...	Silicon hydrogen
SiO_2	...	Silicon dioxide
$SiON$...	Silicon oxynitride
X_{it}	...	Hydrogen species

Chapter 1

Introduction

1.1 Historical Background

In 1926 Julius Edgar Lilienfeld first described a device similar to what we now call a field effect transistor in the US-patent named “Method and apparatus for controlling electric current” [1]. However, it took about thirty more years until the first transistor was actually built; ironically, it was a bipolar junction transistor.

While the first integrated circuits only contained a few transistors, the demand for more complex circuits, and therefore a higher number of transistors, increased steadily. To accomplish the growing number, scaling became the most important topic. In 1974 Dennard *et al.* presented a paper where they stipulated that scaling all device dimensions and voltages by a factor of s at the same time requires to scale all doping concentrations by a factor of $1/s$ to maintain the same electric fields inside the device [2].

In the beginning of the 1980s, the complementary metal-oxide-semiconductor (CMOS) technology was introduced to maintain the development of the already “very large-scaled integration” (VLSI) of transistors. Besides decreasing device size, cleaner and larger fabrication plants for semiconductor manufacture (fabs) were required to increase the yield.

As the demand for faster central processing units (CPUs), larger memory cells, and other integrated circuits increased further, reliability issues concerning the product specifications became more important. In order to reduce the rate of failure of devices further, the semiconductor industry had to improve the involved production processes which often included the replacement of materials responsible for the malfunction of devices. Unfortunately the novel materials in turn caused reliability challenges. One of these reliability phenomena was originally discovered in 1966, when Miura *et al.* linked the generation of charge due to an electrochemical reaction to the presence of a strong electric field at the Si-SiO₂ interface [3], a phenomenon called bias temperature instability (BTI). Despite this, BTI was nearly forgotten for some decades due to its only minor relevance for the early semiconductor industry.

Already right from the start, it was discovered that when interfacing different materials with different lattice parameters, like Si and SiO₂, defects maybe generated at the interface [3,4]. This is due to the non-abrupt transition, which spans over one to two atomic layers and results in an “interface region”, where a lot of dangling bonds act as traps for electrons and holes. By annealing of the structure with hydrogen (H-passivation) the density of these dangling bonds at the interface D_{it} can be reduced from $10^{12} \text{ cm}^{-2} \text{ eV}^{-1}$ to around $10^{10} \text{ cm}^{-2} \text{ eV}^{-1}$ [5], which is a huge improvement. When placing a metal gate electrode on top of the SiO₂-oxide, a metal-oxide-semiconductor (MOS)

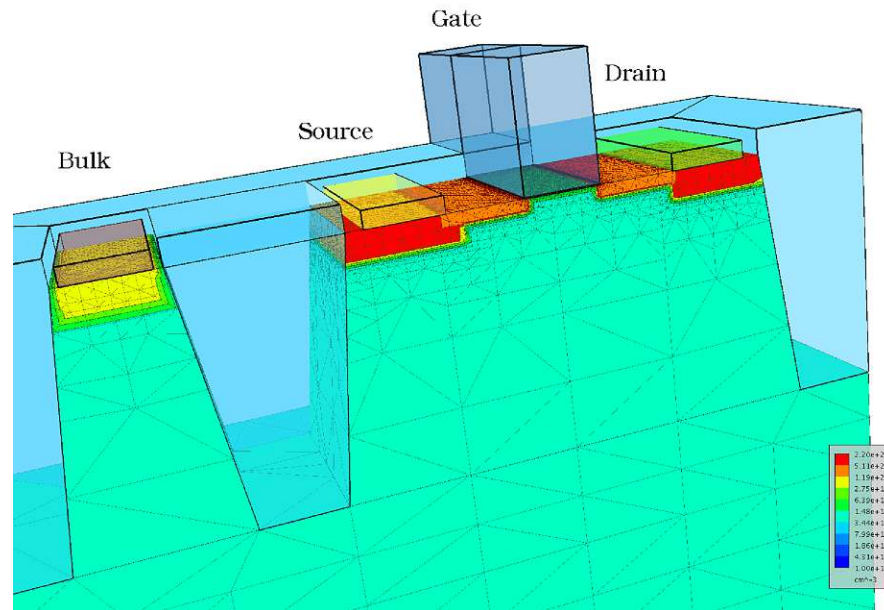


Figure 1.1: An oxide material placed between a gate contact out of metal (aluminum) or highly doped polysilicon and a semiconductor substrate is called metal-oxide-semiconductor (MOS) structure. With ever smaller MOS-structures' as part of the metal-oxide-semiconductor field effect transistor (MOSFET) reliability issues become more important. Note that the doping of the substrate is illustrated on a logarithmic scale with emphasis on the lightly doped drain (LDD) regions between gate and the source, respectively drain regions.

structure is formed, whose operation is explained in Appendix B. Such a MOS-structure is the central part in the metal-oxide-semiconductor field effect transistor (MOSFET), which is exemplarily shown in Fig. 1.1.

As already mentioned, newer materials entered the MOSFET-structure and especially the gate oxide. With the introduction of nitrogen into the oxide the permittivity was increased and the boron diffusion from the gate material into the bulk semiconductor was significantly reduced. At the same time BTI increased in importance¹.

1.2 BTI – Causes and Impacts

The focus of this thesis lies on the advanced characterization of the bias temperature instability and the interpretation of performed stress and relaxation measurements. To be able to understand how BTI affects the MOS-structure, the phenomenon has to be specified first.

BTI happens when the gate of a heated MOSFET is heavily biased while keeping the other contacts grounded [6, 7]. Under these conditions the threshold voltage V_{TH} , the channel mobility μ_{chan} , the transconductance g_m or subthreshold slope, amongst other transistor parameters were shown to degrade.

The most prominent form of BTI when dealing with modern CMOS technologies occurs when the gate of a pMOSFET is biased negatively (in the strong inversion regime); this is called NBTI. When

¹Whether the unexpectedly increased sensitivity to NBTI was due to the considerably increased concentration of defects in the oxide and at the interface or not still remains unclear.

the gate is biased positively, the phenomenon is called PBTI. Including the nMOSFET there are four different permutations of BTI to be distinguished: NBTI/pMOS, PBTI/pMOS, NBTI/nMOS, PBTI/nMOS. Besides the already mentioned case of NBTI/pMOS which exhibits the most dominant effect within the BTI-family, also nMOSFETs show non-negligible PBTI behavior, especially when using high- κ dielectrics. The remaining PBTI/pMOS and NBTI/nMOS combinations are less prone to degrade due to BTI.

As a consequence of BTI, the overall change of the degrading parameters increases the probability that the device fails to meet the specification requirements [8,9], which may yield a malfunctioning device (though not necessarily destroyed yet). Therefore BTI is of industrial as well as scientific interest.

Although silicon as bulk material is a very good heat dissipator to cool the active area inside the MOSFET, the down-scaling mentioned in Chapter 1.1 leads to increasing operation temperatures inside the devices. This increasing operation temperatures slowly move towards the typical NBTI stress temperature ranging between room temperature and 200°C. Due to the increased thermal budget the use conditions for MOSFETs become more demanding.

Unfortunately, at some point during miniaturization the validity of the ideal scaling rule [2] was limited by other factors [10]: Since the on/off current ratio of the MOSFET has to be large enough to be able to distinguish between the signal, the threshold voltage must not be reduced too much. Also, the gate oxide thickness is limited to at least a few atomic layers (≈ 1 nm). These two limitations violate the condition that the oxide electric field E_{ox} remains constant when scaling further; the typically occurring E_{ox} during the operation a MOSFET starts to increase and NBTI becomes more important. Also, tunneling through the oxide and other quantum-mechanical effects become relevant.

During BTI stress the oxide electric field E_{ox} is nearly homogeneous along the channel and thus the description of E_{ox} can be reduced to the vertical oxide electric field $E_{\text{ox}}^{\text{ver}}$ ranging between ± 4 MV/cm up to ± 8 MV/cm and being perpendicular to the interface between oxide and substrate.

When adding a lateral field $E_{\text{ox}}^{\text{lat}}$ by also applying a voltage between source and drain, the carrier velocity in the channel at the drain side increases rapidly. The resulting hot carrier injection (HCI) is supposed to be related to the BTI phenomenon, at least to some extent, but even more complex because of the two electric field components adding up. Hence, a profound knowledge of BTI is required to also understand HCI.

1.3 Modeling BTI with Defects

According to the current understanding BTI is either due to the creation of interface states at the Si-SiO₂ interface, generally known as P_{b} -centers, and/or to the trapping of positive charge inside the SiO₂, possibly at so-called E' -centers. In order to link the observed degradation to these two types of defects as possible underlying physical origin of the BTI phenomenon, the definition

$$\Delta V_{\text{TH}} = -\frac{\Delta Q_{\text{ot}} + \Delta Q_{\text{it}}}{C_{\text{ox}}} \quad (1.1)$$

with Q_{ot} as oxide charge, $Q_{\text{it}} = q_0 N_{\text{it}} f(V_{\text{G}})$ as the charge stored in the interface traps, and N_{it} as the interface trap density and f as the occupancy function at the interface is used. Unfortunately, these charges, whose sum is directly related to the change in V_{TH} , cannot be measured directly. Also, the relative contribution of Q_{ot} and Q_{it} does not necessarily have to be constant over V_{G} .

Nevertheless, till now the defects causing Q_{ot} and Q_{it} are assumed to be the most likely explanation to why BTI happens.

So far, a large number of recent publications have tried to explain NBTI, and in recent years also PBTI. Numerous methods have been devised to classify the contributing defects, their number, spatial positions and energies, and time response. Also, the possible mechanisms of defect creation were thoroughly studied, but up to now, no consensus has been achieved and the debate on the underlying physics and its consequences continues.

Chapter 2

Measurement Methods

To be able to characterize the reliability phenomenon of the negative and positive bias temperature instability (NBTI/PBTI), the experimental access to the degrading and as well observable transistor parameters has to be explained first. State-of-the-art measurements reveal that degradation starts earlier than $1\ \mu\text{s}$ [11] and continues to proceed even beyond weeks [12]. As such, both the onset and the saturation of degradation are outside the experimental window, which today spans about 12 decades in time. The minimum times in this window are due to the limited resolution of the measurement equipment, while the maximum times are restricted by the time a reliability engineer has to perform these kind of measurements¹. Now, a fundamental prerequisite for the description of NBTI lies in an accurate determination of its impacts on the device. But precise measurements of the electric parameters as proper measures of the “real” degradation (e.g. interface state density) are not trivial. This is on one hand due to the immediate relaxation of the degradation once the stress is interrupted, i.e. V_G is set to weak inversion or even accumulation. In 1977 Jeppson *et al.* already described that traps created during negative bias temperature stress can be removed by thermal annealing. The higher the temperature during the annealing process, the quicker the degradation process recovers and the damage is annealed [13]. Nevertheless the NBTI community appeared not interested in the fact that degradation may be reversible under certain conditions for many years. Hence, there was no apparent need to quickly measure the degradation, which of course had a serious impact on the initial modeling attempts. Rangan *et al.* was one of the first to revive the discussion on the recovery of NBTI [14]. A few years later Reisinger *et al.* described the influence of very fast to very slow components contributing to degradation and recovery due to NBTI and contrasted their results to existing physical models in [15], which will be thoroughly discussed in Chapter 3. Today the scientific community has accepted that fast measurements are necessary, but unfortunately there is always a trade-off between a fast and simultaneously accurate method.

This chapter will give a brief overview of the various measurement methods, their delay times, their effect on the device itself, and their other limitations. Moreover, their output signal post-processing complexity is discussed using approximate formulae.

¹As product life cycles are getting ever shorter, accelerated stress tests help the semiconductor industry to determine the data-sheet conditions and error margins of the product, and to finally predict the lifetime of the product.

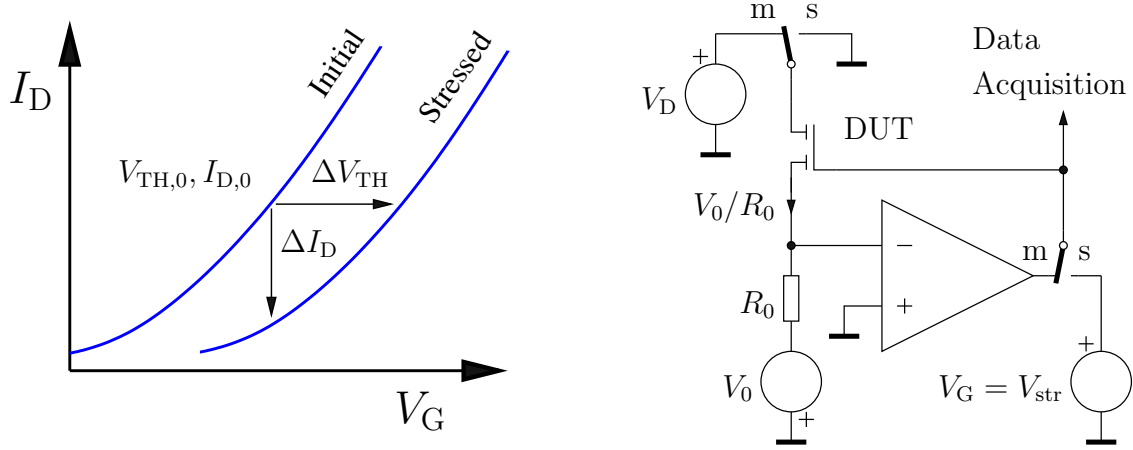


Figure 2.1: **Left:** Schematic picture of an $I_D(V_G)$ -curve before and after stress. The resulting degradation is usually given in terms of ΔV_{TH} or ΔI_D . **Right:** Fast- V_{TH} -method after Reisinger *et al.* [11]. When switched to the measurement-mode ‘m’ the drain current of the device under test (DUT) is forced to a constant (V_0/R_0) by the feedback loop of the operating amplifier. At the same time the threshold voltage V_{TH} is measured. When switched to the stress-mode ‘s’, source and drain are grounded and only the gate is set to $V_G = V_{str}$. The switching between the two modes is done by fast electronic switches.

2.1 Measurement-Stress-Measurement

The probably most widespread measurement technique when investigating BTI issues is the so-called Measurement-Stress-Measurement (MSM) method. In the most simple way the transistor is first characterized through static $I_D(V_G)$ -measurements to obtain a reference of the threshold voltage. (Other ways to extract the onset of the threshold region are listed in [16].) Then V_G is set to V_{str} for some specified time, referred to as the stress time t_{str} . After the end of stress the device is once again characterized and the amount of degradation is estimated by the difference with respect to the initial characteristic. The MSM-method can be performed by either monitoring I_D and a subsequent conversion into a V_{TH} -shift or by directly monitoring V_{TH} .

2.1.1 Monitoring I_D at V_{TH}

One way to assess the NBTI degradation has been suggested by Kaczer *et al.* [17,18], who switch V_G close to the threshold voltage V_{TH} after stress and at the same time monitor the degraded and now recovering drain current I_D over time. By approximating the initial and the degraded $I_D(V_G)$ -curve with quadratic polynomials and assuming that the degradation does not change the form of the initial polynomial approximation, one obtains

$$AV_{TH,0}^2 + BV_{TH,0} + C_1 = I_{D,0}, \quad (2.1)$$

$$A(V_{TH,0} + \Delta V_{TH})^2 + B(V_{TH,0} + \Delta V_{TH}) + C_2 = I_{D,0}. \quad (2.2)$$

Equating these two yields

$$A\Delta V_{TH}^2 + (2AV_{TH,0} + B)\Delta V_{TH} + \underbrace{C_2 - C_1}_{-\Delta I_D} = 0, \quad (2.3)$$

and solving the quadratic form of ΔV_{TH} leads to

$$\Delta V_{\text{TH},12} = \frac{-(2AV_{\text{TH},0} + B) \pm \sqrt{(2AV_{\text{TH},0} + B)^2 + 4A\Delta I_{\text{D}}}}{2A}. \quad (2.4)$$

Using (2.1) and adding I_{D} on both sides yields

$$-AV_{\text{TH},0}^2 - BV_{\text{TH},0} - C_1 + I_{\text{D}} = \underbrace{-I_{\text{D},0} + I_{\text{D}}}_{\Delta I_{\text{D}}}. \quad (2.5)$$

Inserting (2.5) into (2.4) finally gives a formula which only depends on $V_{\text{TH},0}$ and I_{D} . The ΔV_{TH} -shift to the right, respectively the decreasing ΔI_{D} is displayed in Fig. 2.1 left.

$$\Delta V_{\text{TH}} = \frac{-(2AV_{\text{TH},0} + B) \pm \sqrt{B^2 - 4AC_1 + 4AI_{\text{D}}}}{2A} \quad (2.6)$$

This measurement method is generally performed using standard off-the-shelf instruments. Due to the fact that this equipment is not targeted for time-critical measurements, the shortest achievable measurement delays t_{M} only reach down to about 1 ms.

2.1.2 Direct Monitoring of V_{TH}

To improve the measurement resolution of 1 ms, Reisinger *et al.* developed a fast V_{TH} -method [11], which is depicted in Fig. 2.1 (right). It distinguishes two modes of operation: During the measurement-mode a constant and device-specific drain current V_0/R_0 serves as “threshold current” $I_{\text{D}}(V_{\text{TH}})$ -criterium.² This is achieved by a feedback loop using an operating amplifier. Simultaneously, the resulting corresponding threshold voltage V_{TH} of the device is recorded. (The initial reference $V_{\text{TH},0}$ has to be measured in advance.) When switching to the stress-mode all contacts but the gate are grounded, the latter being set to $V_{\text{G}} = V_{\text{str}}$.

With the fast- V_{TH} -method a measurement delay of $t_{\text{M}} = 1 \mu\text{s}$ has been achieved, equivalent to the settling time of the feedback loop. Compared to the studies of Rangan *et al.* [14], who only use off-the-shelf equipment, this results in a three decades faster read-out speed.

2.1.3 Extended-Measurement-Stress-Measurement Setup

To save on time and devices when performing NBTI experiments the extended-MSM (eMSM) measurement routine was established [18]. Choosing each stress sequence $t_{\text{str},i+1}$ to be significantly longer than the previous stress sequence $t_{\text{str},i}$ ensures that the amount of degradation lost during the recovery within $t_{\text{rel},i}$ is nearly completely restored within $t_{\text{str},i+1}$. Consequently, regardless if the stress is interrupted or not, more or less the same amount of degradation is obtained after the total stress time, i.e. $\Delta V_{\text{TH}}(t_{\text{str}}) \approx \Delta V_{\text{TH}}(\sum_i t_{\text{str},i})$. This is schematically depicted in Fig. 2.2, where the top dotted black line of the continuous degradation is always met by the individual sub-sequences (red dotted lines) of the eMSM-sequence after sufficiently long stresses. When the stress sequences are recorded via the on-the-fly method, which will be explained in Chapter 2.3, both stress and recovery can be monitored with the eMSM routine.

²The necessary V_0/R_0 -ratio has to be adapted to the device geometry after $V_0/R_0 = 70 \text{ nA} \cdot W/L$.

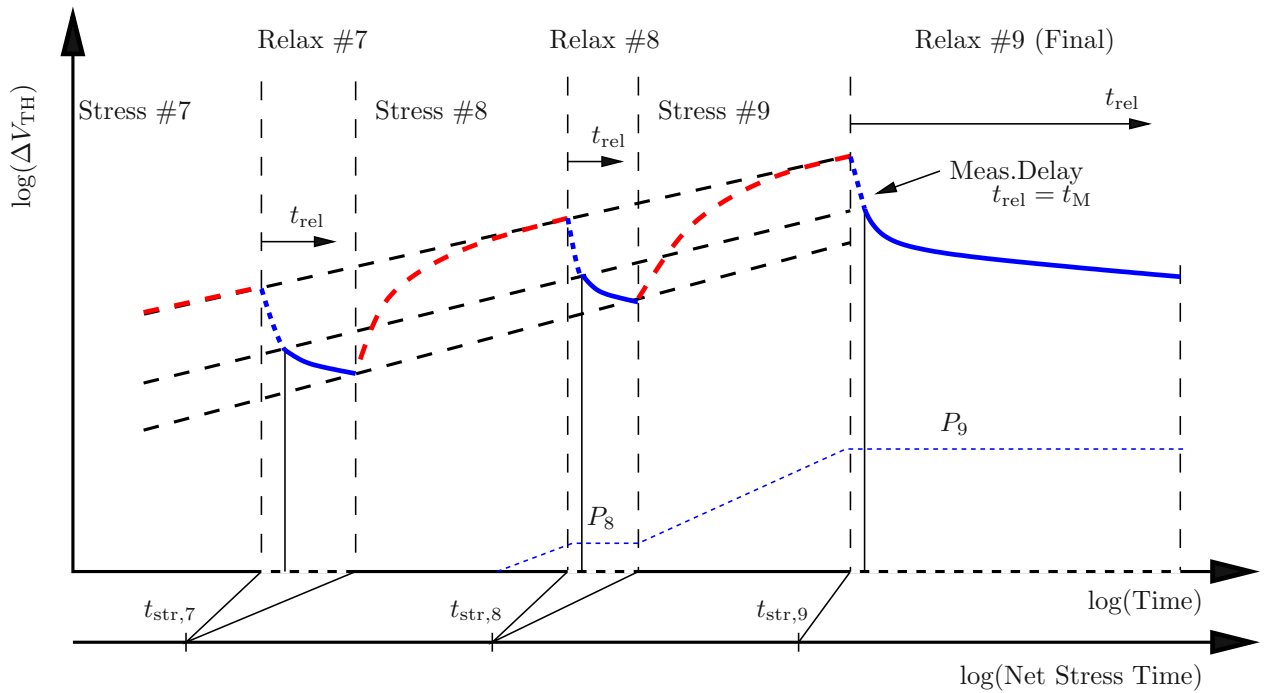


Figure 2.2: Schematic view of the last three out of $N = 9$ stress/relaxation cycles building up an eMSM-sequence like performed by [11,17]. The stress (dashed red) is interrupted $N - 1$ times to record $N - 1$ short and one long final relaxation sequence on the relative time scales $t_{rel} = t - t_{str,i}$. After the measurement delay t_M marked by the dashed blue lines the monitorable relaxation (solid blue) sets in. A permanent or slowly relaxing component P is indicated for the last two cycles and will be explained in Chapter 4.

2.2 Transfer-Characteristics

When carrying out transfer-characteristics measurements, I_D is measured as a function of V_G . Due to their slow response, semiconductor parameter analyzers do not capture the recovery of the device prior to t_M [15], which is schematically depicted by a dashed blue line in Fig. 2.2. Despite their rather slow response ($t_M = 1 \text{ ms} \dots 1 \text{ s}$) parameter analyzers are often used because of their high accuracy.

2.2.1 Fast Pulsed $I_D(V_G)$ -characteristics

Kerber *et al.* [19] were the first to circumvent the problem of slow response times by developing the fast pulsed $I_D(V_G)$ -method shown in Fig. 2.3 (top). They adapted the MSM-technique and used a digital storage oscilloscope (DSO) to quickly measure the voltages and currents of the device under test (DUT) and a programmable pulse-pattern generator. The basic principle of the fast pulsed $I_D(V_G)$ -method is depicted in Fig. 2.3 for NBTI (middle) and PBTI (bottom) and works as follows: During initialization, stress or relaxation, V_G is set to the corresponding constant values V_{rel} , V_{str} or V_{rel} . The pulse generator triggers the fast $I_D(V_G)$ -measurement by sending a gate-pulse reaching from accumulation to inversion when in relaxation-mode, respectively from inversion to accumulation when in stress-mode.

Since the DSO can only measure voltages, the actual drain current is calculated via the voltage drop across R_0 (Fig. 2.3 (top)), assuming V_D small enough so that the transistor stays in the ohmic region.

With standard equipment, pulse times between $100 \mu\text{s}$ [19], $1 \mu\text{s}$ [20–25], down to 100 ns [26] can be achieved. The form of used pulses varies from trapezoidal [19, 21, 26], over rectangular with only very small rise and fall times compared to the pulse width itself [21], up to triangular [20–22, 24, 25]. By varying the rise and fall times of the pulses the trapping and detrapping kinetics can be analyzed [21]. To avoid spurious hystereses (parasitic capacitances) in the $I_D(V_G)$ -characteristics between the rising and falling edges of the pulses, the cable length has to be adjusted in order to ensure the synchronized signal transmission to the DSO [20, 25].

The major issue with this method is that the gain in speed is partly consumed by the fact that the resolution of the DSO is too limited for real ‘single’-pulse-measurements [12]. After the necessary averaging of a few ($10 \dots 1000$) pulses, the measurement time increases by the averaging factor. Furthermore, the synchronization between the pulse-pattern generator and the DSO turns out to be tricky.

2.2.2 Improved Method of Reisinger

State-of-the-art equipment does not meet the combined resolution and measurement speed requirements of NBTI assessment. Instruments either meet (and exceed) the required accuracy, but are too slow to capture the fast NBTI degradation transients (e.g. parameter analyzers), or deliver the necessary time resolution, but are limited by their inherent coarse amplitude resolution (e.g. digital storage oscilloscopes, DSO). Since in the latter case the amplitude resolution can be enhanced by averaging, while in the former there is no remedy for a too slow measurement, a DSO is used to record multiple stress/relaxation-cycles and take the average of these. Care has to be taken to conform to the preconditions of proper averaging, namely to record the *same* process many times. Only in this way, the measurement noise is reduced, while the ‘hidden’ deterministic process is reproduced without introducing systematic errors. In the measurements this is provided by very short stress

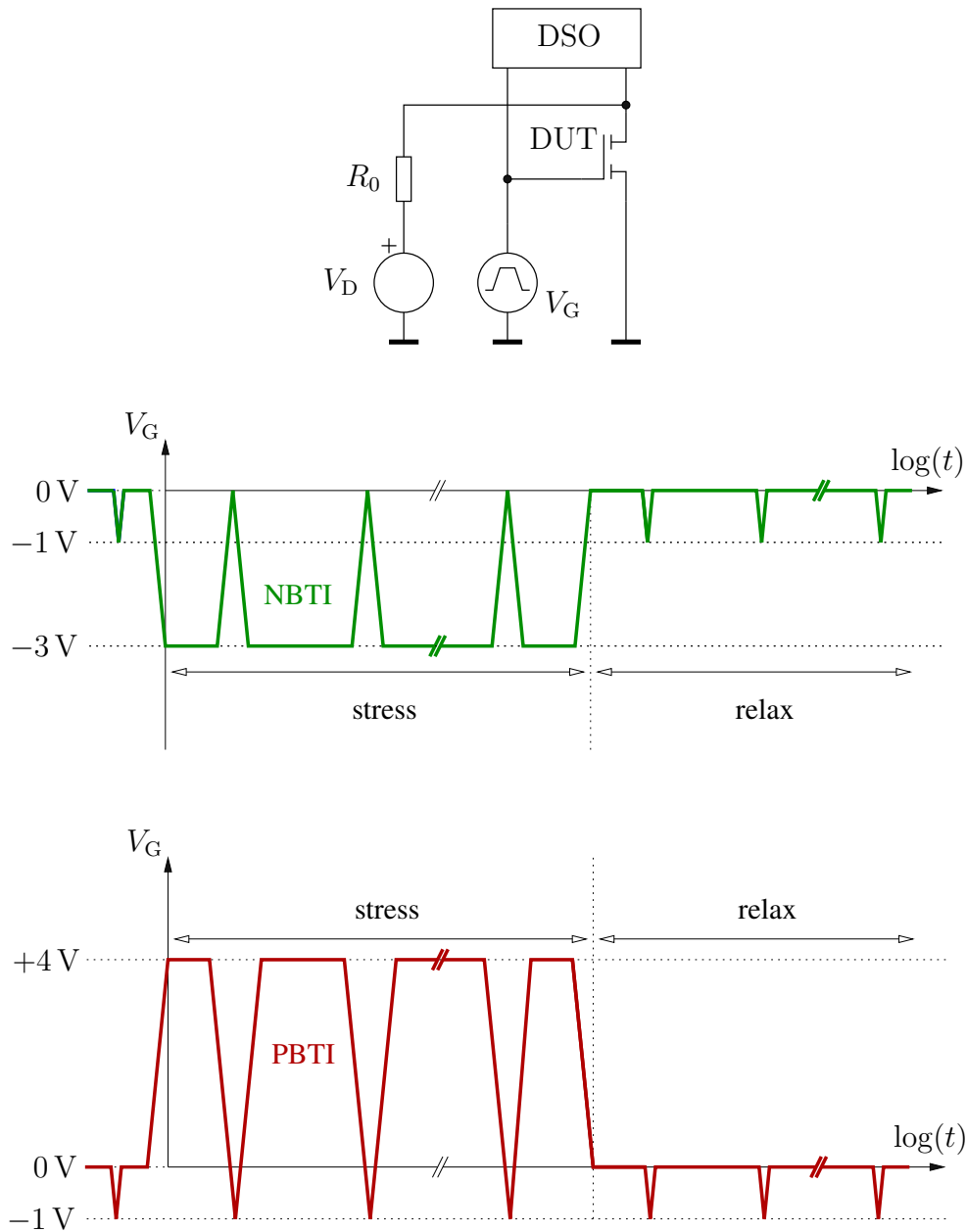


Figure 2.3: **Top:** Kerber's setup to simultaneously record I_D and V_G . Since the digital storage oscilloscope (DSO) can not measure V_D directly, I_D is calculated via the voltage drop across R_0 to finally obtain the $I_D(V_G)$ -characteristic. The corresponding triangular V_G -pulses shown in **Center** and **Bottom** are supplied by the pulse generator. **Center:** The fast-pulsed- $I_D(V_G)$ -characteristics are performed via a superposition of a constant gate level (stress or relaxation) with triangular gate pulses. Switching from the requested NBTI stress of -3 V into the measurement mode ranging from 0 V to -1 V should be carried out as fast as possible in order to avoid undesired relaxation defects. **Bottom:** The same sequence for a PBTI stress of 4 V.

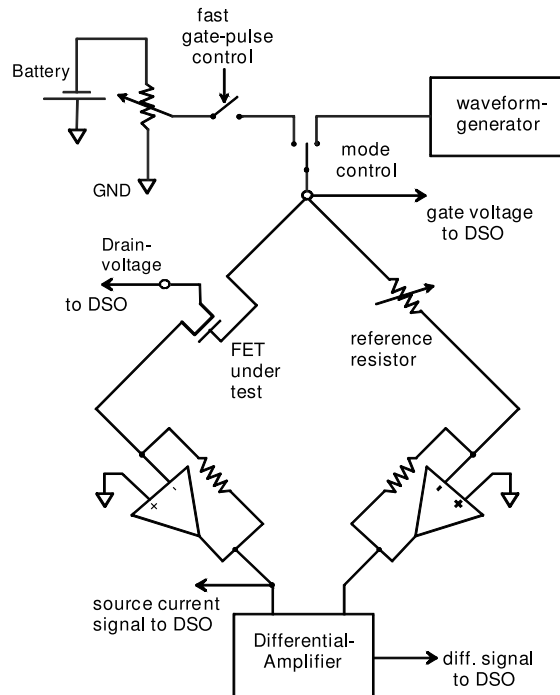


Figure 2.4: Based on the method of Kerber *et al.* the improved fast pulse method by Reisinger *et al.* allows to conduct ultra-short stress measurements. The setup was designed with a bridge circuit containing two differential amplifiers to enhance the signal-to-noise ratio. Since neither commercial voltage sources nor pulse generators were able to fulfill the required settling specification, batteries using a passive voltage divider and fast electronic switches are used in the circuit.

times, and a very low duty cycle in order to achieve nearly 100% relaxation in-between stresses; the characterization due to such a measurement yields a V_{TH} -shift of less than 1 mV.

The method developed in [12] and shown in Fig. 2.4 is related to the previous work of Kerber *et al.* and Shen *et al.* [19, 27] and also used a pulse generator and a digital storage oscilloscope but is able to perform even shorter stress measurements than the previously mentioned methods. Reisinger *et al.* conceived a bridge circuit containing two differential amplifiers. To suppress the noise the I_D of the device under test (DUT) is compared to a reference current, giving only differences, which can then be captured with higher resolution. To furthermore obtain the required resolution of better than 10^{-4} in I_D , the equipment was designed to deliver a settled gate stress voltage $V_{G, str}$ within ± 1 mV in $1 \mu s$. For this reason, a battery using a passive voltage divider and a fast electronic switch are used.

2.3 On-The-Fly (OTF)

While the MSM-technique was conceived to capture the recovery following stress as fast as possible, a completely different approach was first proposed by Denais *et al.* [28]. In contrast to the discussion of the impact of fast recovery which cannot be determined prior to the measurement delay³, the

³Note that the present-day measurement window is reaching down to the μs -regime which seems to be not enough to capture the full characteristics of the recovery after [29, 30].

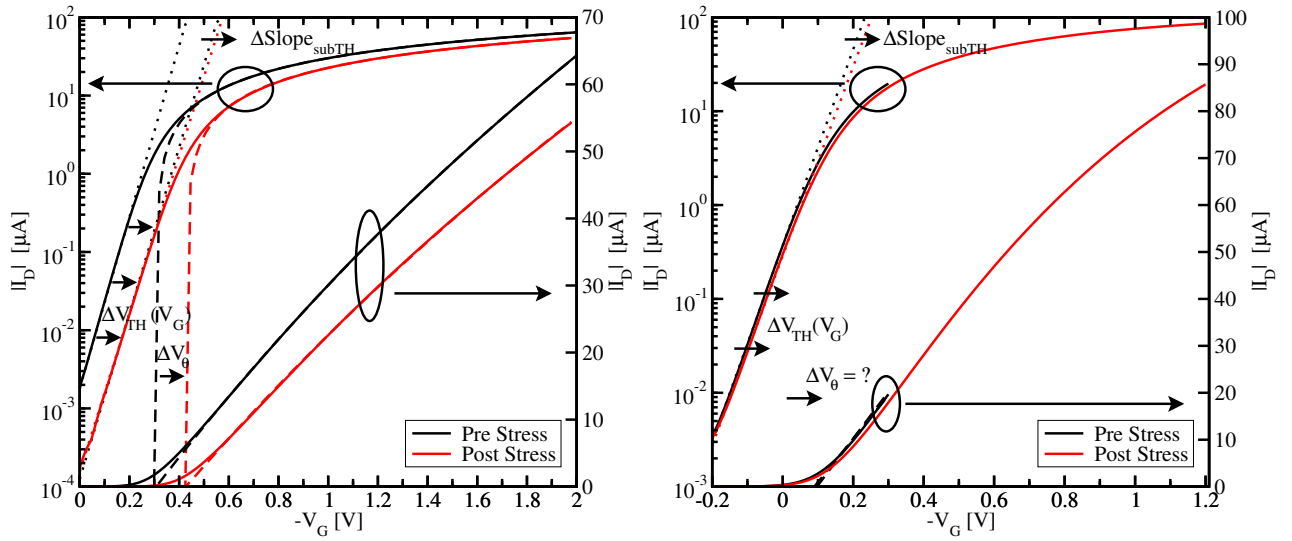


Figure 2.5: Transfer characteristics plotted on a logarithmic- (left ordinate) and linear-scale (right ordinate). **Left:** After stress the $I_D(V_G)$ -characteristics is shifted to the right. The change in the subthreshold-slope due to the increased interface state density affects the physically defined threshold voltage shift, which depends on the gate voltage, i.e. $\Delta V_{TH}(V_G)$. On the other hand ΔV_θ is an empirical quantity, as defined in (2.7). Note that V_θ is larger than $V_{TH}(V_G)$ in the subthreshold regime. **Right:** In contrast to the ΔV_{TH} -extraction in the subthreshold-regime, ΔV_θ has to be determined under strong inversion. Lowering the extrapolation range of V_G decreases the possibility of already pre-stressing the device, but causes an inaccuracy in the thereby determined ΔV_θ .

“on-the-fly” method measures the drain current at stress level without ever interrupting the stress. Due to the experimental setup of never allowing the device to reach the subthreshold regime during stress, the degradation during stress can only be monitored via the degradation of the linear drain current $I_{D,lin}$ [6, 12, 14, 28, 31–34]. Therefore, a method has to be found to convert this measured quantity into a parameter relevant at use-condition, e.g. V_{TH} .

As mentioned in [12], the main problem of the OTF method is that the V_{TH} -shift has almost the same effect on the transfer characteristic as the degradation of the mobility. A shift of V_{TH} as a consequence of electrically active defect charges results in a pure vertical shift along the V_G -axis. More precisely this is because defect charges have a direct impact on the surface potential and hence on the threshold voltage (cf. equation (1.1)). On the other hand, defects located at the interface cause surface scattering. The thereby increased channel resistance (lower mobility) yields a lower drain current after stress and tilts the transfer characteristics. The resulting decrease in I_D than leads to a spurious increase of V_{TH} , in addition to the already mentioned V_{TH} -shift due to the total defect charge itself. Unfortunately, these two effects cannot be separated easily in the linear regime, as can be seen in Fig. 2.5 (left). Due to the saturation of the drain current I_D a relative change in I_D becomes more and more insensitive to changes in V_{TH} with increasing V_G .

The degradation of ΔV_{TH} as defined in (1.1) is just attributed to the defect charges and is independent of the mobility. In contrast to that, $I_{D,lin}$ recorded via the OTF technique does depend on μ_{eff} [34–36], just as it reflects the existence of additional charges (ΔQ_{ot} and ΔQ_{it}). To extract

$I_{D,\text{lin}}$ the simple SPICE compact model [37] valid in the linear regime under strong inversion only is used:

$$I_{D,\text{lin}} = \frac{\beta V_D (V_G - V_\theta - V_D/2)}{1 + \theta (V_G - V_\theta - V_D/2)} \quad \text{for} \quad V_G > V_\theta. \quad (2.7)$$

While β depends on μ_{eff} , θ models the mobility saturation with increasing vertical field and V_θ , the threshold voltage, is obtained by the intersection of $I_{D,\text{lin}}$ extrapolated to $I_{D,\text{lin}} = 0$, which is depicted in Fig. 2.5 (left). Due to the fact that the interface charge depends on the gate voltage through the occupancy at the interface, as stated in (1.1), the threshold voltage is not a well defined quantity, i.e. $\Delta V_{\text{TH}} = \Delta V_{\text{TH}}(V_G)$ [37, 38]. Equation (1.1) uses a physical definition of a threshold voltage, while V_θ is a purely empirical quantity that yields the best fit to the level 1 model⁴. It can be shown that it is important to provide a large V_G -range to get a reliable extraction of V_θ .

The main issue with OTF is that as a matter of principle it is not possible to determine the initial $I_{D,\text{lin}}$ at $t_{\text{str}} = 0$, because due to the nonzero measurement time the device is already stressed, and so the first measurement yields $I_{D,\text{lin}}(t_{\text{str}} > 0)$. This pre-stressed value is then taken as a reference, which has a considerable impact on the subsequent extraction of the degradation [39–41].

When the V_G -range is reduced as depicted in Fig. 2.5 (right), at least for the pre-stressed transfer-characteristic, a value close to the initial value, i.e. $I_{D,\text{lin}}(t_{\text{str}} \approx 0)$ is obtained. On the other hand this method induces a large error, which is of the same order of magnitude as ΔV_θ itself. Therefore, it is not feasible to describe the $I_{D,\text{lin}}$ -regime properly by reducing the V_G -range.

Different OTF models are based on (2.7) and are discussed in Appendix A in detail. Here the so-called OTF3 after Zhang *et al.* [34], displayed in Fig. 2.6, will be described. A change in $I_{D,\text{lin}}$ can only be converted to ΔV_θ if the transconductance g_m , which is defined as the change of the I_D over V_G , is known. To get g_m , I_D is recorded while slightly varying V_G . This three-point measurement method [28] is indicated in Fig. 2.6 as well and yields

$$g_m(n) = \frac{I_{D,\text{lin}}(V_G + \Delta V) - I_{D,\text{lin}}(V_G - \Delta V)}{2 \Delta V}. \quad (2.8)$$

By averaging g_m , ΔV_θ is finally obtained via the sum

$$\Delta V_\theta^{\text{OTF},3} \approx - \sum_{n=1}^N \frac{I_{D,\text{lin}}(n) - I_{D,\text{lin}}(n-1)}{1/2 (g_m(n) + g_m(n-1))}. \quad (2.9)$$

In order to prevent a degraded reference of $I_{D,\text{lin}}$ and g_m , Zhang *et al.* suggested to perform the oscillation of V_G with a rise and fall time of $6 \mu\text{s}$. Considering such a “degradation-free” reference thus produces a higher amount of visible $\Delta V_\theta^{\text{OTF},3}$ -degradation [42] due to the down-shifted initial value of $I_{D,\text{lin}}$ and g_m . Moreover, as ΔV_{TH} increases with V_G , the OTF-method measures a higher degradation ($\Delta V_\theta(I_{D,\text{lin}})$) compared to the typical use-condition of a device ($|V_{G,\text{use}}| < |V_{G,\text{str}}|$). OTF hence overestimates the “real” degradation. In contrast the “real” degradation is underestimated, when the evaluation of V_{TH} is based on DC transfer characteristics. As a consequence, the determination of the lifetime is heavily influenced by either measurement routine. Datasheet conditions on the other hand should better reflect the real degradation under real use-conditions of devices.

Compared to MSM, the biggest advantage of OTF is its recovery-free measurement routine while it is difficult to measure recovery with it, because the OTF technique originally was conceived only to record data in the stress phase of NBTI.

⁴The question whether ΔV_{TH} or ΔV_θ should be preferred will not be discussed. Usually circuit-designers use ΔV_θ while physicists prefer ΔV_{TH} .

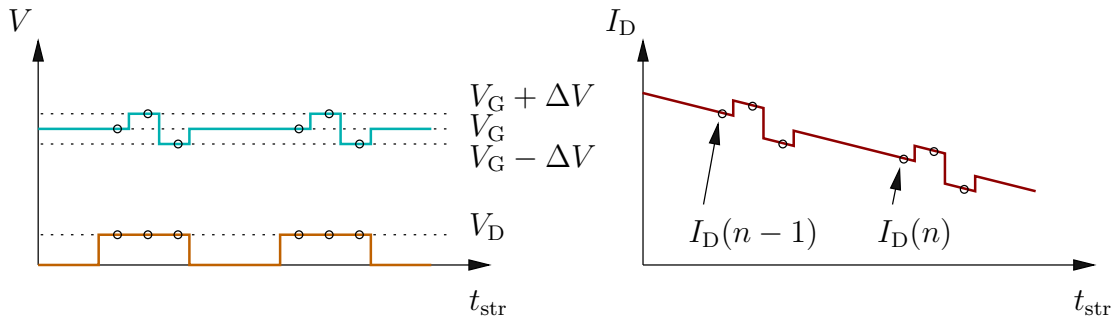


Figure 2.6: Schematic of the OTF3 methodology. **Left:** The three points symbolize the quantities at V_G and their small perturbation $\pm\Delta V$. The drain voltage V_D stays constant during the pulse. **Right:** The resulting $I_{D,\text{lin}}$ whose two points $I_D(n-1)$ and $I_D(n)$ are needed to determine the degradation of I_D . The shift of g_m is calculated via (2.8) by using the modulated $I_{D,\text{lin}}(V_G)$.

2.4 Charge Pumping

For the assessment of interface traps, which are situated at the interface between the oxide and the substrate along the channel of a MOSFET, the charge pumping method (CP) is often the measurement method of choice. The setup of CP after [43, 44] is described on basis of Fig. 2.7 (left). Source and drain of the transistor are shortened and biased at a certain reverse voltage with respect to the substrate. To perform a CP measurement the gate is stepped between accumulation and inversion by a pulse generator. Pulsing towards inversion deeply depletes the surface and minority carriers (holes in the case of the depicted pMOS) are injected from the source and drain regions into the channel where they can be captured by the interface states. When going back from inversion towards accumulation, the mobile minority carriers drift back to source and drain due to the reverse bias. The charges trapped at the interface are too slow to follow, but will recombine with the majority carriers of the substrate (electrons). The resulting recombination current, the charge pumping current I_{cp} , is measured at the bulk. I_{cp} is proportional to the number of interface traps, i.e. CP directly measures the amount of interface states.

By additionally varying parameters like amplitude, frequency, rise and fall time of the pulses and the temperature of the device under test, not only the number of interface traps, but also their energetic and spatial distributions can be determined, which gives very useful information of NBTI related degradation [45–49].

Unfortunately, the characterization by CP measurements following NBTI stress, especially the characterization of the fast recovery behavior, is extremely challenging because the CP technique inherently relies on a bias switch into accumulation. Consequently, it is unclear whether the often observed weak recovery in CP data is a consequence of the fact that interface states do not recover or whether this is an artifact of the measurement technique brought about by the strong bias switch. Moreover, the necessary averaging of many pulses implies a rather large measurement delay, which is also not favorable when trying to access early recovery.

2.5 On-the-Fly Fast Charge Pumping

In order to avoid possible recovery effects, the on-the-fly charge pumping technique, also called on-the-fly fast interface trap (OFIT) technique was developed by Li *et al.* [24, 25, 51]. As illustrated

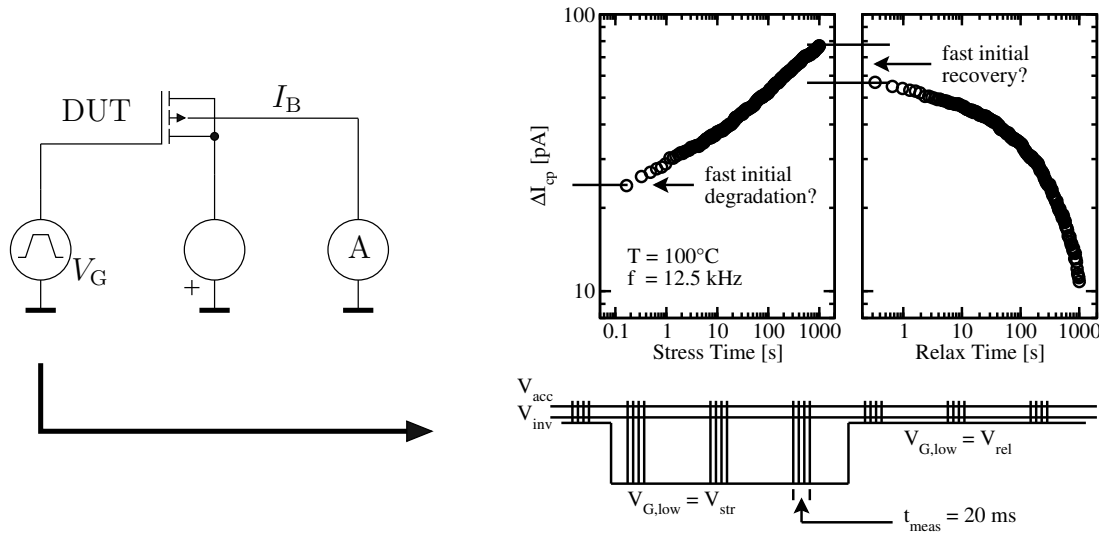


Figure 2.7: **Left:** Setup of the charge pumping measurement technique to indirectly determine the amount of interface states. A gate pulse sweep is applied ranging from accumulation to inversion. Thereby, the minority carriers that got trapped by the interface states during inversion recombine with the majority carriers during accumulation. This recombination current is then proportional to the interface state amount. **Right:** Typical NBTI stress and relaxation measurement of the charge pumping current I_{cp} using the OFIT technique with a duty cycle of 50%. Each symbol in the upper graph consists of some hundred averaged pulses as schematically displayed in the lower graph for some points. After the reference measurement, where $V_{G,low} = V_{rel}$ is pulsed across V_{acc} and back periodically, $V_{G,low}$ is set to the stress level V_{str} , in order to continuously stress the device. In this way the CP measurement and the application of stress are carried out consecutively with only interrupting the stress by the very short gate pulse. During relaxation $V_{G,low} = V_{rel}$. Constant slopes of the stress and relaxation pulses have to be ensured [50].

in Fig. 2.7, the basic difference between OFIT and CP is that the low-level $V_{G,low}$ of the CP pulse is simultaneously used as a stress condition (for NBTI), while the actual CP measurement is performed by quickly switching back and forth between accumulation V_{acc} and stress V_{str} . Consequently, as will be discussed in Chapter 5, the low-levels are different during stress and recovery/reference measurements, which is also depicted in Fig. 2.7. To decrease the recovery further, a low duty cycle is chosen.

2.6 Capacitance Voltage Profiling

Since in a semiconductor the local carrier density depends on the position of the Fermi level, and hence on the local electrostatic potential, even a simple parallel plate capacitor structure with one plate replaced by a bulk semiconductor (MOS) exhibits strongly non-linear behavior [10, 52]. The charge distribution of minority and majority carriers in such a MOS structure varies as a function of the applied gate voltage. Depending on this charge distribution, different operating regimes can be identified. When going from accumulation towards inversion, the depletion region inside the semiconductor is formed first. Here, the majority carriers are driven away from the interface between the oxide and the semiconductor. The only remaining charges within this depletion region are fixed ionized acceptors (p-type) or donors (n-type), which build up a depletion charge. In combination with the insulator this results in a decrease of the total capacitance.

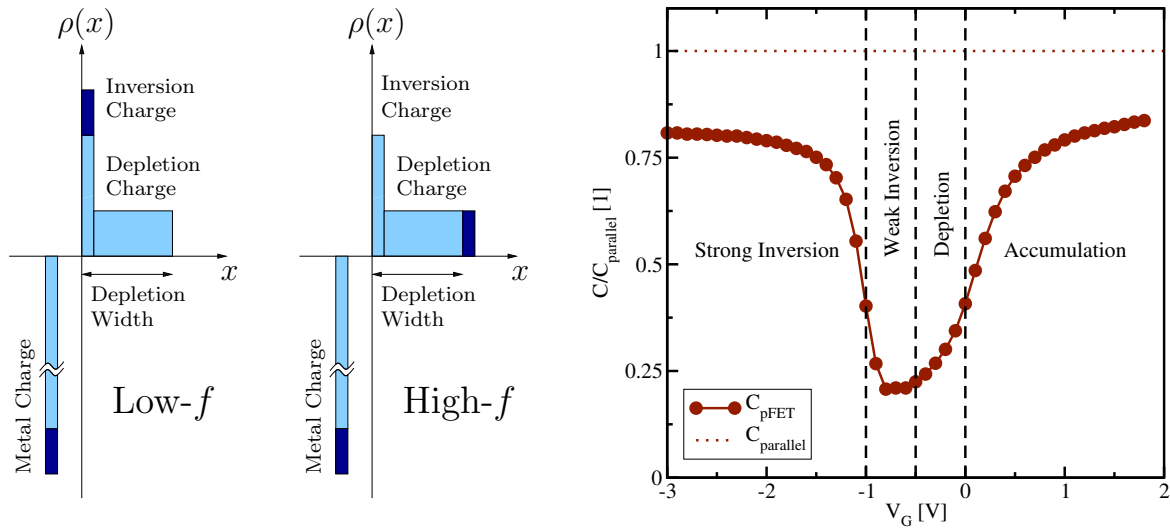


Figure 2.8: **Left:** The frequency dependence of the MOS-structure capacitance. While for low frequencies the minority carriers in the inversion layer are fast enough to contribute to the signal, the increase in charge can only be compensated by an increase of the depletion width (majority carriers) at higher frequencies. **Right:** A $C(V)$ -curve of a pMOSFET taken at 1 MHz and normalized to the capacitance of an ideal capacitor of the same size. To be able to distinguish between the different operating regimes more easily, the graph is also shifted by the workfunction difference. The flatband capacitance C_{fb} is then obtained at $V_G = 0$ V.

With the onset of inversion the minority carriers exceed the majority carriers at the interface and create the inversion layer. At that point the depletion region with its ionized impurities virtually stops to increase in width and any increase in gate charge is only balanced by an increase of the inversion charge. Whether the minority carriers are able to follow the signal or not influences the capacitance in this inversion regime, i.e. the contribution of the inversion layer charge to the total capacitance depends on the frequency. Only at low frequencies the recombination-generation rates of the minority carriers can keep up with small signal variations leading to a charge exchange with the inversion layer. With this additional inversion charge the capacitance signal increases as the depletion width remains constant [10]. At high frequencies on the other hand only the majority carrier response is measured. Hence the incremental charge in deeper inversion is put at the edge of the depletion region, while the inversion regime is not altered. This causes the capacitance to remain constant when going from depletion into inversion. Both cases (low and high frequency) are depicted in Fig. 2.8 (left).

When adding source and drain regions to form a MOSFET, minority carriers are provided independently of the frequency. Therefore low-frequency $C(V)$ -characteristics of MOSCAPs and $C(V)$ -characteristics of MOSFETs look alike. Exemplarily, a $C(V)$ -characteristics of a pMOSFET measured at 1 MHz is depicted in Fig. 2.8 (right). For a better understanding the curve shown is shifted by the flatband voltage⁵ and the different operation regimes of the MOSFET are marked.

Based on the above mentioned findings the $C(V)$ -characteristic provides valuable information of the semiconductor structure and its interface. For example, present interface states stretch the $C(V)$ -characteristic along the V_G -axis, because additional charge is necessary to fill these traps.

⁵The flatband voltage results from the workfunction difference of the materials used.

Oxide charges on the other hand are independent of the applied V_G and cause a mere parallel shift of the $C(V)$ -characteristic towards higher or lower V_G [10]. Furthermore, with the knowledge of the capacitance as a function of V_G , the oxide electric field can be calculated. This is necessary when the degradation caused by NBTI is compared with that caused by PBTI for the same device type, e.g. for a pMOS. Due to the nonzero flatband voltage it is not possible to apply just the opposite V_G to achieve the opposite electric field. Moreover, the different behavior of the capacitance during accumulation and inversion yields an asymmetric $C(V)$ -characteristic. This as well influences the value of the proper (exact opposite) field. An application of $C(V)$ -characteristics to obtain the required stress voltage $V_{G,\text{str}}$ for a certain NBTI stress and its corresponding PBTI stress is given in Chapter 7.

Chapter 3

Previous Modeling Attempts

In the 1960's, the investigations of the Si-SiO₂ interface revealed a close-coupling of the increase of surface traps sitting at the interface and a phenomenon which will be later on known as the bias temperature instability. Both effects were known to cause a negative shift of the threshold voltage [3, 4, 53]. Though this phenomenon was already known not to cause real device failure as for example time dependent dielectric breakdown (TDDB) [35, 54], the creeping shift of V_{TH} alerted the industry and the scientific community to develop a model which is capable of describing the mechanisms behind BTI. In order to judge such a model as functional, clear definitions of its applicability but also potential limits have to be listed. The following review summarizes the existing modeling efforts, including their advantages and disadvantages.

3.1 Reaction Diffusion Model

As the fabrication of a semiconductor device today as back then consists of more and more single layer depositions, often being followed by an annealing step, hydrogen as “the” passivation agent was suggested to play a key role in the first modeling attempt dating back to 1977 [13]. In the so-called reaction-diffusion model Jeppson *et al.* assumed the breaking of hydrogen-bonds at the interface via a thermally and field-activated process under stress. This reaction-limited stress phase is schematically depicted in Fig. 3.1 (taken from [55]) and can be described with the kinetic rate equation at the interface after [31, 56–58] as

$$\frac{\partial N_{it}}{\partial t} = k_f(N_0 - N_{it}) - k_r N_{it} X_{it}^{1/a} \quad (3.1)$$

where N_0 denotes the total amount of interface states, N_{it} the fraction of dangling bonds thereof (not yet passivated), and X_{it} the interfacial hydrogen concentration. The rates k_f and k_r describe the forward (depassivation) and reverse (passivation) process with a kinetic exponent a considering the “size” of the diffusing species.

3.1.1 Stress Phase

As long as the bond-breaking dominates the rate equation, the reverse rate is negligible because there is simply not enough free hydrogen. Thus the degradation within this initial stress phase is only proportional to the stress time t_{str} .

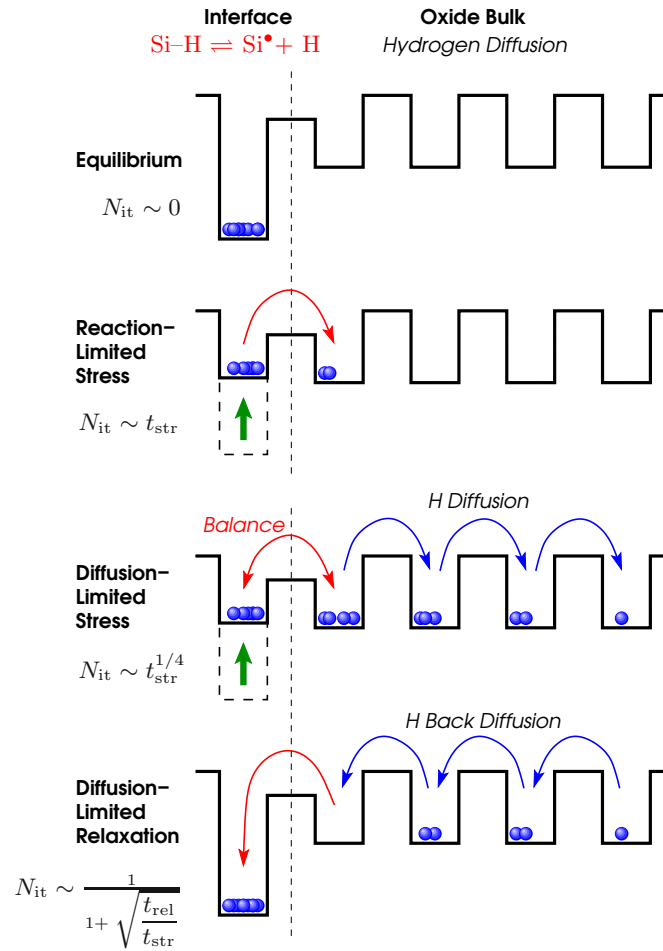


Figure 3.1: From **Top to Bottom**: Schematic view of the atomic hydrogen reaction-diffusion model during stress and relaxation. During the short initial phase, the interface region enters equilibrium. When in equilibrium, the degradation is dominated by the diffusion of hydrogen. As the stress is removed, the hydrogen diffuses back to the interface.

Once the interface reaction reaches equilibrium, the previously released hydrogen species diffuse towards the oxide. Under the assumption of a well passivated interface with only a few initial dangling bonds $N_{it,0}$ and atomic hydrogen H^0 as diffusion species the diffusion-limited stress regime can be approximated by a power-law of the form $t_{str}^{1/4}$. The mathematics behind this diffusion-limited process can be found in Appendix C.

However, not all measurement results are consistent with the predictions of the reaction-diffusion theory using atomic hydrogen. Quite to the contrary, the extracted exponents were found to depend on the measurement delay time; the exponents of 1/4 were obtained for measurement data with large delay time only [7,59]. Shorter delay times on the other hand yielded exponents of around 1/6. Chakravarthi *et al.* now interpreted this weaker time dependence by introducing instant dimerization of the released atomic hydrogen at the interface via $H^0 + H^0 \rightleftharpoons H_2$ and subsequent diffusion of the created H_2 [60]. This theoretical assumption yields a smaller exponent of approximately 1/6 because of the larger kinetic exponent ($a = 2$ for H_2 instead of $a = 1$ for H^0 , cf. Appendix C). When performing even faster measurements with delay times around a micro-second, e.g. OTF-measurements

done by Reisinger *et al.*, exponents of 0.12 were obtained [12] which does not correspond to the stress behavior predicted by the reaction-diffusion theory.

3.1.2 Back Diffusion of Hydrogen during Recovery

The phase after switching off a device after a certain stress time or between switching on and off is called recovery or relaxation. During this recovery phase the degraded parameters start to revert to their initial values, but within times that depend on the prior stressing conditions. Now the question arises how the interplay between BTI stress and recovery during the operation of a MOSFET does affect the reliability of the device. This is important because only when characterising and modeling both stress and recovery a realistic lifetime extrapolation of the device is possible.

Within the RD theory the recovery is explained via back diffusion of hydrogen. Once the stress is removed, the quasi-equilibrium at the interface causes all interfacial hydrogen to immediately passivate the dangling bonds. By the time all interfacial hydrogen is bound and no more hydrogen is available at the interface, hydrogen located deeper in the oxide is assumed to diffuse back and control the entire recovery as the limiting process. During recovery the back diffusion can be approximated by the empirical expression

$$N_{it} \sim \frac{1}{1 + \sqrt{\frac{t_{rel}}{t_{str}}}}. \quad (3.2)$$

This equation is related to the universal recovery [61] which will be discussed in Chapter 4.1 more thoroughly. Interestingly, the diffusion-limited recovery in (3.2) yields about 90 % recovery within 4 decades of time, whereas experimental data still show recovery over at least 10 decades of time [61]. Moreover, since the RD recovery only depends on the ratio of relaxation to stress time, the model as such is not capable of explaining any other experimentally observed recovery behavior, e.g. dependence on temperature or stress voltage. The RD model is also not able to explain the dynamic behavior¹ of NBTI when applying alternating stress and relaxation sequences with a varying duty factor (DF) or duty cycle (DC) [6, 20, 30, 62].

3.2 Extensions of the Reaction-Diffusion Model

The lack of a decent description of recovery based on the reaction-diffusion theory, cf. (3.2), soon urged the development of modified and extended RD models [17, 33, 41, 55, 59, 63–67]. In the following variants thereof are summarized.

1. As modern oxide layers are only a few nanometers thick, the explanation based on diffusing hydrogen inside the oxide was questioned. Therefore the diffusion process of H₂ was suggested to continue inside the polysilicon as well [63]. By assuming two different diffusion coefficients, namely fast diffusion in the oxide and slow diffusion inside the polysilicon, the so-called **two-region RD model** was expected to be able to explain the much larger observed recovery range. Actually this range only increased at little [55].
2. The **two-interface RD model** [33], presented by Krishnan *et al.*, focuses on the quick diffusion of atomic hydrogen inside the oxide. Once having reached the polysilicon interface,

¹The correct dynamic description is of utmost importance as it corresponds to real use conditions.

a second chemical reaction takes place creating molecular hydrogen ($H^0 + Si-H \rightleftharpoons Si^\bullet + H_2$). Like in the two-region model, the molecular hydrogen diffuses further into the polysilicon. Since the diffusivity in the oxide is regarded to be very high compared to the diffusivity in the polysilicon, the H^0 stored in the oxide is indeed able to cause a fast initial recovery. For large stress times, on the other hand, it is this higher oxide diffusivity that locks the hydrogen in the polysilicon for a long time. This means that the short recovery effect vanishes.

3. In contrast to the two-region RD model, where instant dimerization at the interface is assumed, the **RD model with explicit dimerization** is based on a continuous dimerization process inside the oxide, what allows both hydrogen species to coexist while diffusing into the oxide [64]. Whereas the initial stress phase is thereby altered to $t_{str}^{1/3}$, the recovery characteristic remains the same compared to the standard RD model.
4. Since the experimentally observed recovery revealed a log-like characteristics (cf. Section 4.1), Islam *et al.* questioned the interface states to be fast enough to follow the gate voltage V_G switches. They suggested an **RD model assuming slow interface states**. Unfortunately, such a model is in stark contradiction to the Shockley-Read-Hall theory (SRH) used to describe the trapping dynamics at the interface with transients due to electron capture being within the nano-second regime. Under the assumption of excessively small capture cross sections some sort of fast relaxation in the microsecond-regime within one or two decades in time is indeed obtained. However, this form of recovery is not observed in any experimental data [55].
5. **Extended reaction-dispersive-diffusion (RDD) models** using a broad distribution of energy levels were discussed in [17,55,59,65,66]. They describe the hydrogen transport occurring via the highest energetic states only (transport level). Hydrogen being located in a deeper energy level needs to be thermally activated prior to be able to diffuse further into or out of the oxide, i.e. without any activation this hydrogen is trapped. Further, in these models only hydrogen sitting at the interface is allowed to re-passivate which slows down the reverse rate as most of the hydrogen is trapped.

In contrast, a simplified version of the RDD model does not differentiate between trapped and untrapped hydrogen, i.e. all hydrogen is allowed to interact with the interface [67]. This implies a faster initial recovery, compared to the non-simplified RDD model, cf. simulations performed in [55].

Although with increasing dispersion of the bond breaking at the interface the recovery can be slowed down, none of the RDD variants is finally able to describe the actual experiment.

The following conclusion can be drawn for RD theory in general. While during recovery solely passivation occurs, the stress is modeled using depassivation and passivation simultaneously [31]. At present, no extension of the RD-model is able to describe recovery after stress in a reasonable form. Whether such a model is then able to describe the much more complex stress-relaxation patterns during the operation of a MOSFET is very questionable. The premises are simply not correct. This leads to the conclusion that hydrogen diffusion is very unlikely to be a main player when dealing with NBTI degradation. For this reason completely new approaches are inevitable [6, 11, 18, 40, 61, 68, 69].

3.2.1 Dispersive-Reaction-Rate Models

Due to the amorphous structure of the interface, the binding energy of the Si-H-bonds at the interface is not constant but varies from site to site. Electron-spin-resonance (ESR) studies revealed

the binding energies as distributed Gaussian with a variance of 0.02–0.08 eV [70], i.e. weaker bonds break first, while stronger bonds remain passivated. Longer stress times or a larger applied electric field are required to break those stronger bonds [71].

Charge pumping (CP) as the measurement of choice for the assessment of the amount of interface states (cf. Chapter 2.4) revealed some interesting facts. The observed amount of recovering interface states accessed via CP after NBTI stress was too small to be able to explain the overall recovery of ΔV_{TH} . Therefore a part of the community [62, 72–74] considered the generated interface states as permanent once created. This assumption will be discussed in Chapter 8.

Quite in contrast, Mahapatra *et al.* stated that CP measurements in the range of seconds are too slow to detect the recovery of interface states because of inherent delay of the measurement setup. Another possibility to explain the missing recovery involves the CP technique itself, as it pulses into accumulation which in turn causes unwanted additional relaxation [75].

In order to clarify the issue of how interface states contribute to recovery, Li *et al.* developed the on-the-fly fast interface trap CP method (OFIT) [25, 51], described in Chapter 2.5. Based on the results of this OFIT method [24, 51, 76], which showed recovery faster than a second, but also revealed long-term recovery, Grasser *et al.* derived a BTI-model based on interface states only in [77]. Therein they describe two distinct components of the recovery as two facets of a single degradation mechanism proceeding as a series of steps. By assuming dissociation of Si–H bonds (dispersive bond breaking) the so-called double-well (and subsequently refined triple-well) model is able to describe quite complex stress-relaxation-patterns. Though the mathematics in this model describe the NBTI phenomenon correctly, its microscopic assumptions are likely unjustified [78], an issue that will be examined in detail in Chapter 5.

Chapter 4

Two Components Contributing to Bias Temperature Instability

As of today, no model can successfully explain all peculiarities of the BTI phenomenon. While many groups have already rejected the approach using diffusing hydrogen five years ago [6, 11, 30], Mahapatra *et al.* still keep the RD theory alive for “predicting NBTI stress and recovery” in [79]. However, a model explaining BTI requires the understanding of its contributing mechanisms which are not necessarily straightforward. According to the RD theory the time dependence of the threshold voltage during BTI can be modeled by using forward and backward rates. During stress these forward and backward processes were assumed to take place simultaneously, implying a superposition of both processes. During recovery on the other hand the forward process was supposed to vanish. Due to this circumstance the recovery was considered to be the key issue which has to be studied first before dealing with the more complicated stress phase. Therefore precise and commonly practiced stressing¹ and relaxation routines are more than helpful. Note that the measurement techniques presented in Chapter 2 do not interpret BTI degradation and relaxation the same way. In fact, the obtained measure of the degradation/relaxation is often monitored using different equipment with varying delay times. Furthermore, different types of quantities are obtained from the different measurement techniques, be it the linear drain current $I_{D,lin}$ [28] using on-the-fly characterization, the ΔV_{TH} -shift at a certain drain current [11] using the measure-stress-measure approach (MSM), or an $I_D(V_G)$ -measurement to extract V_{TH} using a digital storage oscilloscope (DSO) [20]. Hence, a detailed consideration of the proper measurement setup and procedure is of utmost importance to be able to accurately determine the real degradation, at least as far as possible.

Concerning the mechanisms causing this degradation, the scientific opinion is divided whether hole trapping is insignificant and only the interface traps degrade and recover [31], or whether trapped oxide charges on top of the creation of interface defects are relevant [6, 40, 49, 66, 81]. In 2006 Huard *et al.* first stated the existence of a recoverable component R on top of a permanent component P [6]. To be able to understand which kind of microscopic mechanism or defect forms which component, another very important attribute of the BTI recovery has to be explained first.

¹Not only repeatability but also a certain standard procedure to be able to compare measurement results for modeling perspectives is feasible. Such arrangements are discussed in the JEDEC meetings that fix the global standards for the microelectronics industry on a regular basis [80].

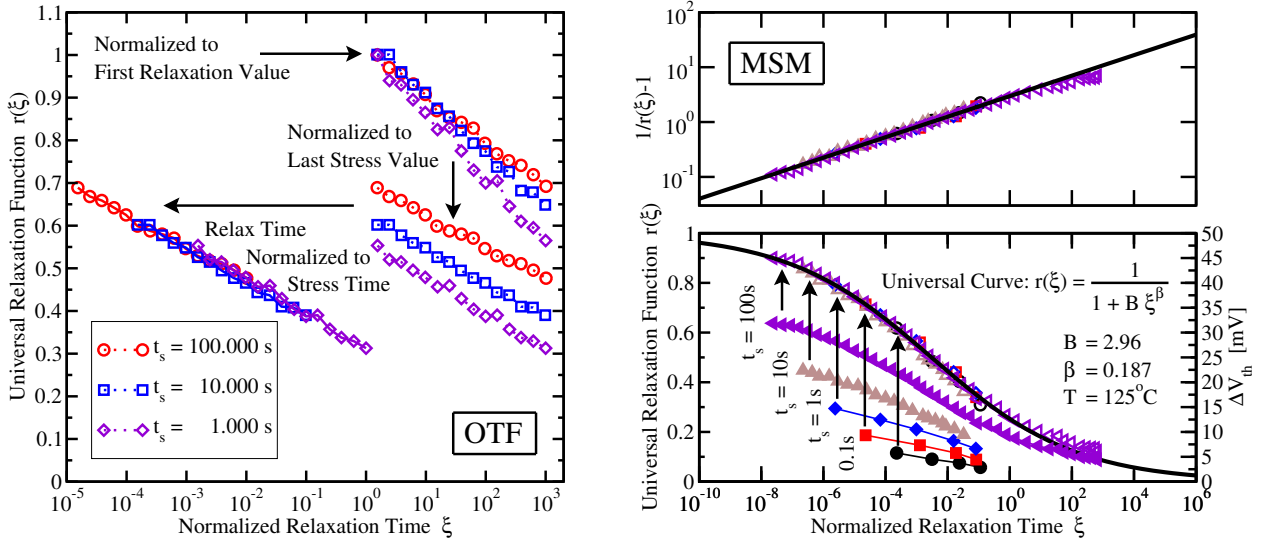


Figure 4.1: **Left:** Demonstration of how universal relaxation works for OTF data of Denais *et al.* [28]. Starting with the data normalized to the first relaxation value, the second step is to refer the relaxation curves to the last stress values. By then dividing their relaxation time by their corresponding stress time yields perfect universality for all three stress times. **Right:** When fitting the MSM data of [11] by (4.3), many relaxation traces are required to solve for B and β since the first recovery point $R(t_{\text{rel}} = 0)$ is unknown due to the measurement delay. The linear behavior is depicted by $1/r(\xi) - 1$ on a log/log plot (**Top Right**). Note the slight deviation for larger ξ , which is due to an existing permanent component and in fact makes the correction of the universal curve (4.2) unavoidable.

4.1 Universality of BTI recovery

It has been shown that the reaction-diffusion theory is not able to predict the different observed recovery characteristics following from different stress conditions. Therefore scaling and normalization routines are empirically used to characterize the recovery traces [14, 33, 61, 82].

In Fig. 4.1 (left) the normalization routine is depicted step by step. First, the relative recovery² after different $t_{\text{str},i}$ is normalized to the last corresponding stress value. Furthermore, all relaxation times are divided by the last stress times, yielding $\xi = t_{\text{rel}}/t_{\text{str}}$. The normalized data now lie on top of each other and feature the same “universal” behavior. Based on [61], this universal behavior can be described by the universal relaxation function $r(\xi)$ which reads

$$r(\xi) = \frac{R(t_{\text{str}}, t_{\text{rel}})}{S(t_{\text{str}}) - P(t_{\text{str}})} = \frac{R(t_{\text{str}}, t_{\text{rel}})}{R(t_{\text{str}}, 0)}. \quad (4.1)$$

Here $R(t_{\text{str}}, t_{\text{rel}})$ denotes the relaxation depending on the total stress time t_{str} and the relaxation time t_{rel} , $S(t_{\text{str}})$ the total degradation observed right at t_{str} , $P(t_{\text{str}})$ the permanent component, which only depends on the total stress time, and $R(t_{\text{str}}, 0)$ stands for the total amount the device is going to recover from $t_{\text{rel}} = 0$. As such the term universality in NBTI was not a new finding in 2007, but was already reported by Denais *et al.*, who claimed that the relative recovery follows the same pattern when plotted over the stress time $\xi = t_{\text{rel}}/t_{\text{str}}$ [61, 82].

²The relative recovery denotes the absolute recovery normalized to its initial value.

Unfortunately yet no generally accepted and valid theory exists for BTI, which means that the exact form of $r(\xi)$ remains speculative. Empirically derived expressions have therefore been presented so far: Kakalios *et al.* for example used a stretched-exponential of the form $r(\xi) = \exp(-B\xi^\beta)$ to describe “relaxation in disordered systems” [83]. Other empirical “log-like” expressions are $r(\xi) = 1 - \beta \log(1 + B\xi)$ [82] or $r(\xi) = \beta \log(1 + B/\xi)$ [84]. Also a generalized power-law-like expression after [61]

$$r(\xi) = 1/(1 + B\xi^\beta), \quad (4.2)$$

amongst many others [17, 85–87], can be used. Whereas equation (4.2) can be used to fit the OTF-data from [82], the first recovery point $R(t_{\text{rel}} = 0)$ of the MSM-data, depicted in Fig. 4.1 (right), is unknown due to the instant recovery. The first MSM point is obtained with a delay time t_M and according to [11, 61] even a $t_M = 1 \mu\text{s}$ is still too slow. Thus back-extrapolation to reconstruct the “true” initial degradation has been suggested. With $\xi_M = t_M/t_{\text{str}}$ we get

$$\frac{r(\xi)}{r(\xi_M)} = \frac{1 + B\xi_M^\beta}{1 + B\xi^\beta}. \quad (4.3)$$

Provided that there is a set of relaxation data with different stress times, B and β can be determined, as shown in Fig. 4.1 (right) taken from [11, 61]³. The slight deviation for $\xi > 10^2$ is due to the existence of a non-negligible permanent component $P(t_{\text{str}})$, which is additionally present here and consequently has to be considered too.

4.2 Assumption of a Permanent Component

Grasser *et al.* refined the approach of [6, 88] to obtain universality irrespective of the amount of $P(t_{\text{str}})$. In [61] they presented a correction scheme for (4.3), which is based on the determination of the accumulated degradation due to stress after the delay time t_M as

$$S_M(t_{\text{str}}, t_M) = R_M(t_{\text{str}}, t_M) + P(t_{\text{str}}). \quad (4.4)$$

Here, S_M splits up into R_M , the recoverable amount of degradation monitored at t_M , and a permanent components P which is regarded as independent of t_M . In [29], $P(t_{\text{str}})$ was supposed to follow a power-law of the form At_{str}^n . In order to characterize the temperature and bias dependence of the components of (4.4), plasma-nitrided-oxide (PNO) devices with an effective oxide thickness (EOT) of 1.4 nm and 2.2 nm were characterized. Therefore the OTF-method [17] and the fast- V_{TH} -method developed by [11] were used. The latter method is embedded into an eMSM-sequence which is carried out with N stress/relaxation-subsequences, already described in Chapter 2.1. A typical eMSM-measurement is shown in Fig. 4.2 (left).

For the extraction of R and P , the yet unknown permanent contributions of the single relaxation phases P_i have to be determined simultaneously. The remaining non-permanent parts of the relaxation sequences are then fit to the universal relaxation law (4.2). Altogether this yields a relaxation model with $N + 2$ parameters

$$\begin{aligned} S_M(t_{\text{str},i}, t_{\text{rel}}) &= R(t_{\text{str},i}, t_M) \frac{r(t_{\text{rel}}/t_{\text{str},i})}{r(t_M/t_{\text{str},i})} + P_i \\ &= R(t_{\text{str},i}, t_M) \frac{1 + B(t_M/t_{\text{str},i})^\beta}{1 + B(t_{\text{rel}}/t_{\text{str},i})^\beta} + P_i. \end{aligned} \quad (4.5)$$

³Further details of the application of the universal relaxation law towards modeling and the influence of the measurement delay are discussed in [61].

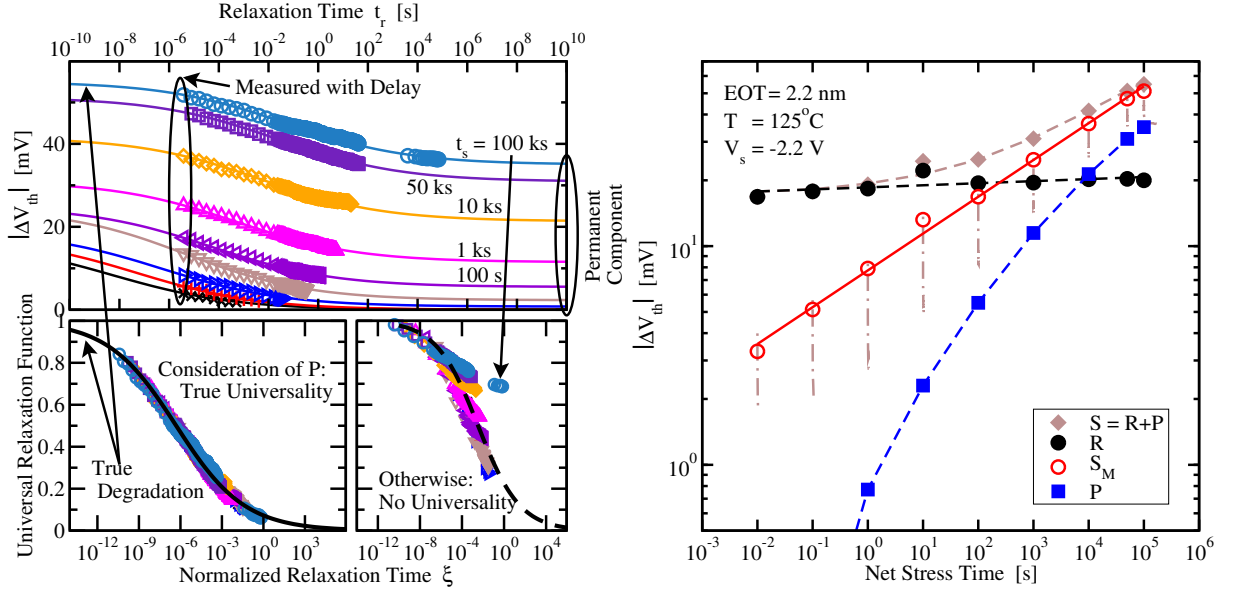


Figure 4.2: **Top Left:** The measurement data (symbols) of an MSM-sequence with very short delay of $1\ \mu\text{s}$ is matched on the relaxation model (4.5) with $N + 2$ parameters. It yields perfect universality over more than 10 decades in time (lines). **Bottom Left:** After subtracting the single P_i of each relaxation sequence (marked above), all data can be fit to a single line. The universality of the relaxing component R is clearly visible. **Bottom Right:** Without considering P_i , data at large stress times does not conform to the universality. **Right:** The original measured data $S_M(t_{\text{str}}, t_M \approx 1\ \mu\text{s})$ and the extracted recoverable and permanent components R and P . By back-extrapolating $S_M(t_{\text{str}}, t_{\text{rel}} = t_M)$ the “real” total degradation $S(t_{\text{str}}, t_{\text{rel}} = 0)$ is obtained, consisting of $R + P$ (compare to the true degradation depicted left). The relaxation data in-between the stress sequences is indicated by dash-dotted lines on a relative time scale $t_{\text{str},i} + t_{\text{rel}}$. The recoverable component R can be well fit by either a power-law or $n \log(1 + At_{\text{str}})$ (used in the following) while P closely follows $P_{\text{max}}/(1 + (t_{\text{str}}/\tau)^{-\alpha})$ after [6].

Therein, B and β are fit parameters for the universal recoverable component R , and the P_i with $i = 1 \dots N$ denote the N relaxation sequences which have to be optimized. The results of the optimization loop are then illustrated in Fig. 4.2 (right), clearly showing the existence of a permanent (or slowly relaxing) component [30], when the recovery levels off. In contrast to R , which can be fitted by a power-law or $n \log(1 + At_{\text{str}})$, P behaves like a power-law for shorter stress times only. It clearly shows signs of saturation at longer stress times, which is fundamental for lifetime extrapolation. Without considering such a permanent component, universality is not given, like shown in Fig. 4.2 (bottom left).

Moreover, it is of utmost importance to study wide relaxation periods, as the data gained that way yields a much more reliable basis for modeling, compared to other measurements done on commercial equipment: While Alam *et al.* covered about 3 decades in time [49, 85], the widest recovery behavior observed with commercial equipment so far accounted for 5 decades time [6, 40, 66]. Using their dedicated equipment Reisinger *et al.* [11] were able to measure BTI relaxation periods of 10 to 12 decades in time with the shortest available delay time of $1\ \mu\text{s}$, cf. Fig. 4.2 (left).

Applying the universality on various pMOS/nMOS-NBTI/PBTI-combinations yields different quantitative, but all in all consistent results. Surprisingly, this also applies to the negative shift of the threshold voltage they all have in common, for details refer to Fig. 4.3.

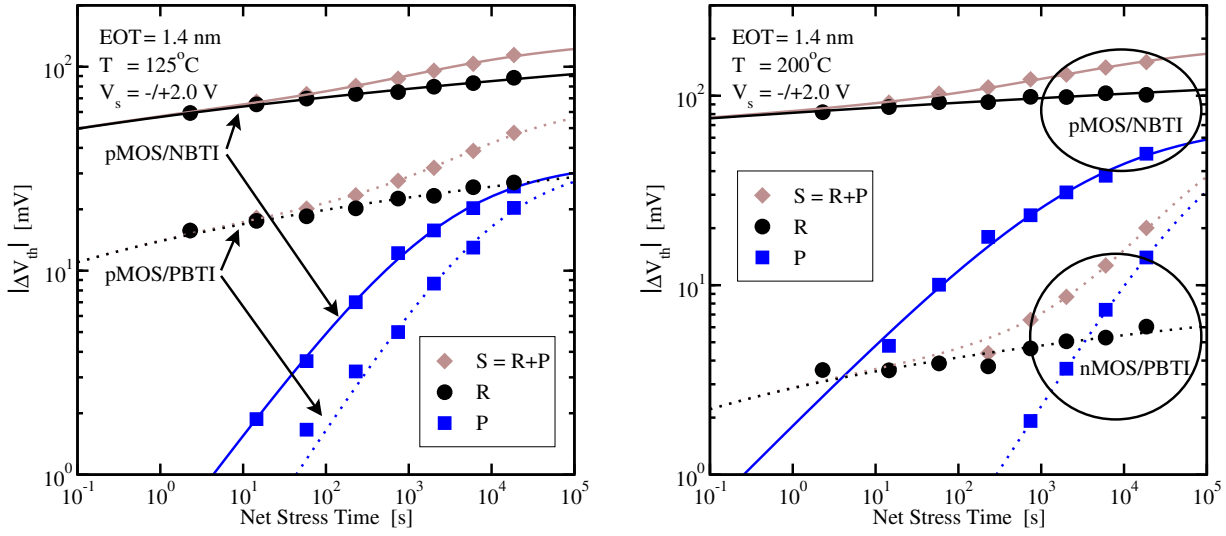


Figure 4.3: The universal relaxation during PBTI/NBTI stress depicted for pMOS and nMOS transistors (symbols: data, lines: model). The samples have a much smaller EOT of 1.4 nm compared to the devices of Fig. 4.2 with 2.2 nm, a considerably larger R but a comparable P component. **Left:** pMOS: The recoverable component during PBTI stress is qualitatively the same but shifted by a factor of 3 to smaller values, while the permanent component is quite similar compared to NBTI. **Right:** Comparison of pMOS/NBTI and nMOS/PBTI stress. The recoverable component of the nMOS is very small (6 mV). Hence, the degradation is dominated by P from an early stage when comparing it to pMOS.

4.2.1 Temperature and Voltage Dependence of Universal Law

Up to now it was only shown that the universality holds for various pMOS/nMOS-NBTI/PBTI-combinations. As temperature and voltage acceleration play an important role for lifetime projection, the study of the universal relaxation is now extended towards these stress conditions. How the two components R and P behave is therefore analyzed under different stress temperatures T and stress voltages V_S , cf. Fig. 4.4 (left). In this graph only the last long relaxation tail of the MSM-sequence is depicted. The different stress conditions described by the relaxation model (4.5) yield excellent agreement with the measurement results. The activation energies E_A are extracted for B and β , and the components of R and P . They are depicted in Fig. 4.4 (right). While R and B show an Arrhenius-like behavior with $E_A \approx 0.08$ eV respectively 0.04 eV for different stress times, β is constant. P on the other hand depends on the stress time, which rules out Arrhenius-like behavior [6, 30].

4.2.2 Measurement Delay

Based on the extracted model parameters of (4.5), a correlation between the observed degradation and the measurement delay can be obtained. The actually observable data marked with S_M in Fig. 4.5 (left) is bound between $S = R + P$ (the extrapolated ‘true’ degradation) and P . The larger the delay time, the closer S_M and P get and vice versa.

When fitting the single stress sequences with varying t_M by a power-law, different values of the slope are obtained which may be a reason why the power-law exponents reported in NBTI literature vary that strongly. In Fig. 4.5 (right) the power-law slopes, defined as $d \log(S_M) / d \log(t_{str})$, are

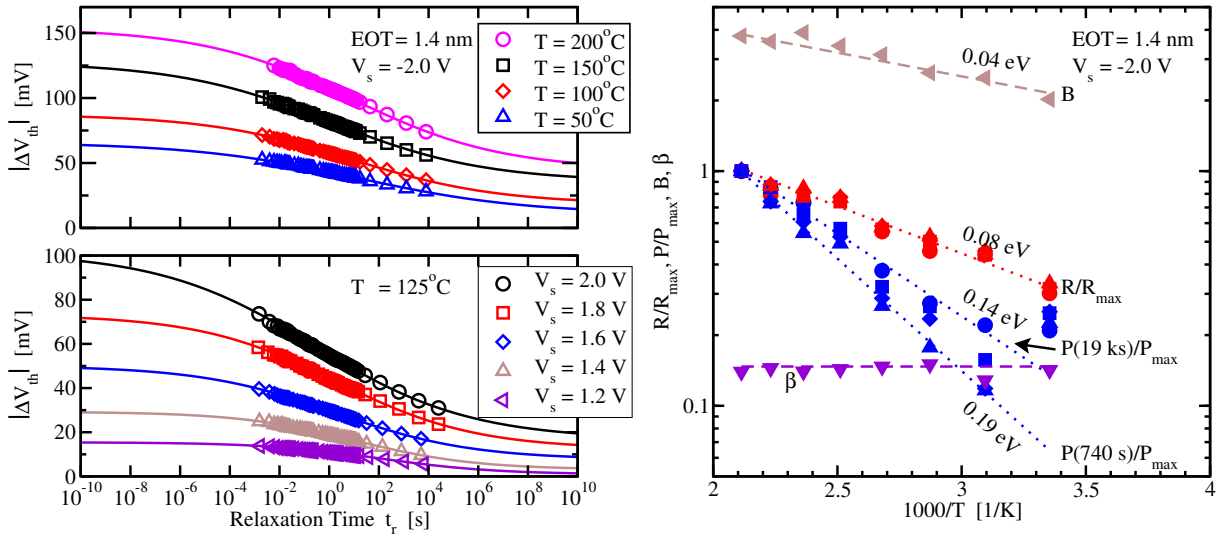


Figure 4.4: **Left:** When only the last relaxation sequence after 20000 s stress is considered under various temperatures (**Top**) and voltages (**Bottom**), the fit shows universal behavior. **Right:** Using different stress times and temperatures allows to extract activation energies in an Arrhenius plot for the four components R , P , B , and β in our relaxation model (4.5). While R and B are Arrhenius, P is not (due to different values of E_A), and β is constant.

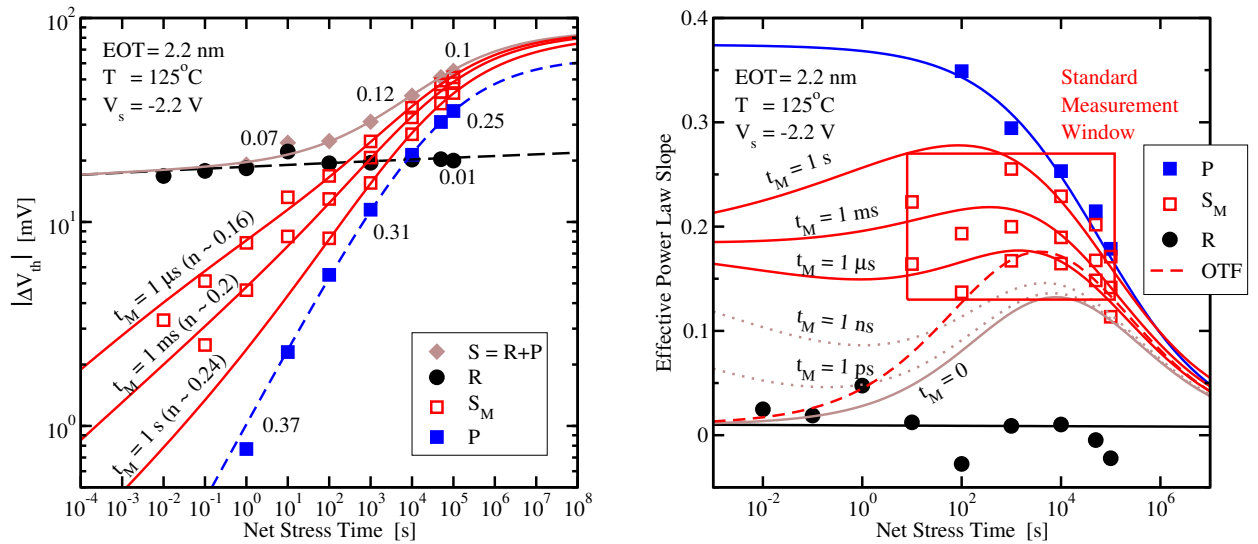


Figure 4.5: Influence of the measurement delay on the observed NBTI behavior (open symbols: data, closed symbols: extracted R and P , lines: model). **Left:** The measurement results (S_M) lie between S and P . Depending on the measurement delay of the equipment (t_M) a broad range of ‘effective’ power-law slopes are observed (limiting values given next to the model lines). **Right:** The effective power-law slope as a function of the measurement delay, defined as $d \log(S_M)/d \log(t_{str})$ is only approximately valid over a few decades in time within the standard measurement window. This is due to the interaction between R (depending on the measurement delay) and P (independent of t_M).

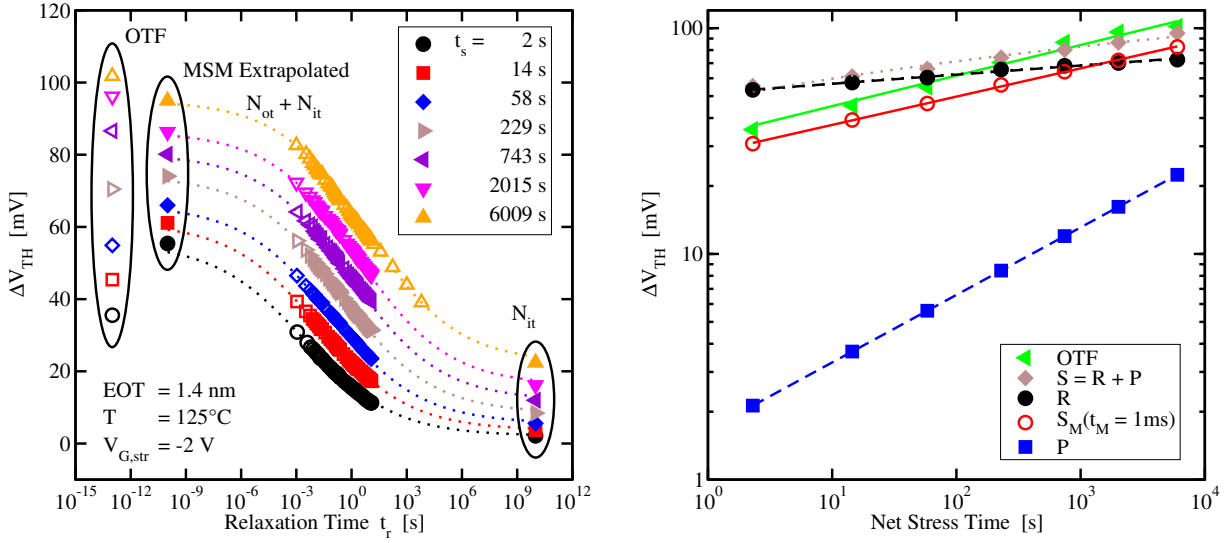


Figure 4.6: Comparison of measured MSM and OTF1 data (open symbols). **Left:** Using the universal relaxation law the different stress sequences can be reconstructed to the very first relaxation moment, i.e. the full NBTI degradation is obtained (closed symbols and lines). The degradation observed using OTF1 measurements for the same devices does not correspond to the MSM relaxation data. **Right:** The OTF1 data, the recoverable component R on top of a permanent/slowly relaxing component P , their sum S , and the pure MSM data evaluated with a delay of $t_M = 1$ ms are extracted from the left figure as a function of the stress time. The overall degradation of $S = R + P$ does only qualitatively agree with the OTF1 data points.

shown. As can be seen the extrapolation with a power-law does not seem to be the best choice to represent the time behavior of NBTI due to the interplay between R and P . Since the power-law extrapolation is furthermore only approximately valid over a few decades in time, lifetime prediction based on this approximate concept should be done with great care.

4.3 ΔV_{TH} versus ΔV_θ

Based on the determination of S , S_M , R , and P , the change of V_{TH} obtained from the measurement-stress-measurement (MSM) routine will now be compared to the change of ΔV_θ obtained from the on-the-fly (OTF) method. The results provide valuable information on the applicability of these two measurement routines and furthermore give new insights into the yet not too well known dynamics of the contributing defect states.

By using the universality (cf. Section 4.1) [30,61], the full degradation of an MSM-stressed device is reconstructed in Fig. 4.6. Unfortunately, the extrapolated initial values right after stress do not match the degradation gained by the OTF1-method. To explain the differences, the numerical device simulator MINIMOS-NT [89] is used. Applying the drift-diffusion transport model after [90], Boltzmann statistics for the carrier concentrations [10], Shockley–Read–Hall (SRH) interface state dynamics after [91], and mobility variation due to interface state Coulomb scattering [92], a well defined number of defects (N_{it} and N_{ot}) is placed at the interface of a pMOS as used in [17]. The simulated I_D is then post-processed the same way, as already done by the MSM-setup and finally converted to ΔV_{TH} . By using definition (1.1) of $\Delta V_{TH} = -(\Delta Q_{ot} + \Delta Q_{it})/C_{ox}$, a

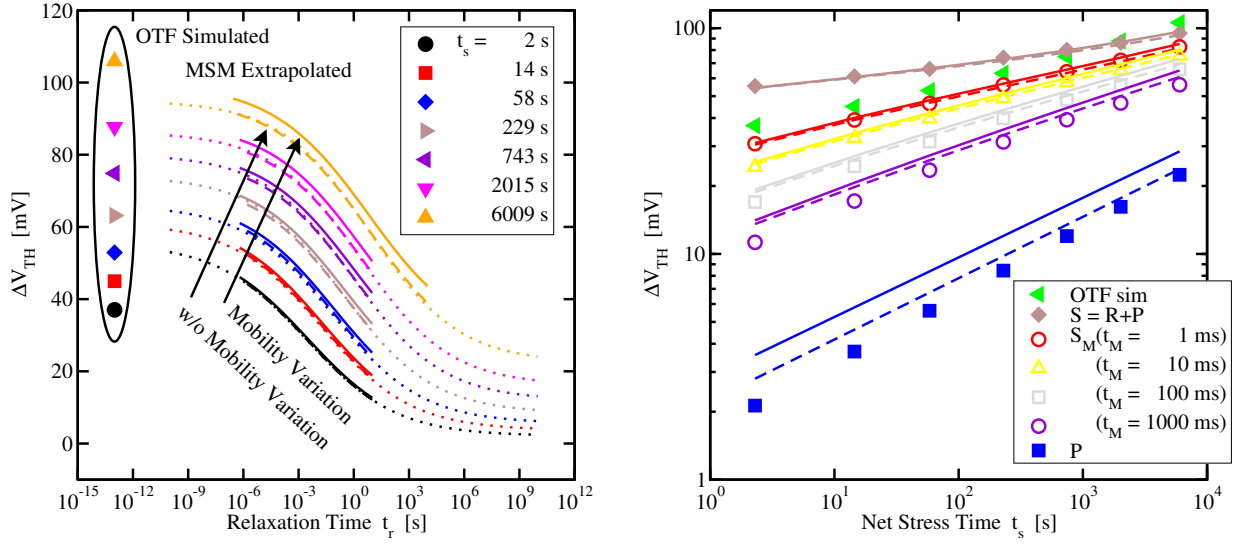


Figure 4.7: The simulations of the MSM-sequence and OTF-methods based on the measurement sequence depicted in Fig. 4.6 yield sound agreement. The degradation is considered to be due to oxide charges (recoverable part) and interface states (permanent part). **Left:** The extracted ΔV_{TH} once including (solid) and once not including (dashed) mobility variation confirms that the extraction only slightly depends on the mobility. The maximum error of 4 mV is observed for $t_{str} = 6000$ s. The dotted lines are taken from Fig. 4.6. **Right:** The simulation results as a function of the stress time (lines) with the calculated $S = R + P$ and P components (closed symbols) fit the measurement data (open symbols). Solid lines are with 10% mobility degradation while dashed lines are without. The faster the MSM-sequences are recorded, i.e. the smaller t_M , the more S_M approaches S .

parametric relationship between ΔV_{TH} and the resulting charge caused by defects is obtained [39]. The occupancy of the interface states $f(V_G)$ determines the detectable charges following the relation $Q_{it} = q_0 N_{it} f(V_G)$, where $f(V_G)$ results in a change of the subthreshold-slope. This finally affects the calculated ΔV_{TH} , as already indicated in Fig. 2.5.

Simulating the MSM-sequence, shown in Fig. 4.7, yields excellent agreement when mobility changes during stress are neglected. However, many publications have emphasized that mobility variations impact the accuracy of the OTF-method [35, 41]. Simulations performed in [39] showed that an estimated error of 3% in the effective mobility results in a spurious shift in ΔV_{θ} of about 50 mV^4 . The error in ΔV_{TH} obtained after an MSM-simulation is roughly 5 mV for the same device which denotes only a tenth of the simulated ΔV_{θ} -shift. Grasser *et al.* confirmed these results as being due to the impact of the mobility variation on the extracted threshold-voltage shift. Obviously this impact depends on the applied gate voltage. By employing a numerically simulated $I_D(V_G)$ -characteristics including a 10% mobility degradation, they showed that the impact of the mobility is largest in the linear OTF-regime and only weak in the subthreshold MSM-regime [39]. Consequently, the determination of ΔV_{TH} should be carried out with V_G safely in the exponential regime of I_D to avoid additional mobility effects.

When now the measurement sequences depicted in Fig. 4.6 are re-simulated with and without a mobility variation of 10% after 100 ks stress, as done in [36], Fig. 4.7 (left) is obtained. The extracted ΔV_{TH} perfectly fits to the expected values given by Q_{it} and Q_{ot} when mobility changes are neglected,

⁴Note that this high value might already be in the range of the device failure definition.

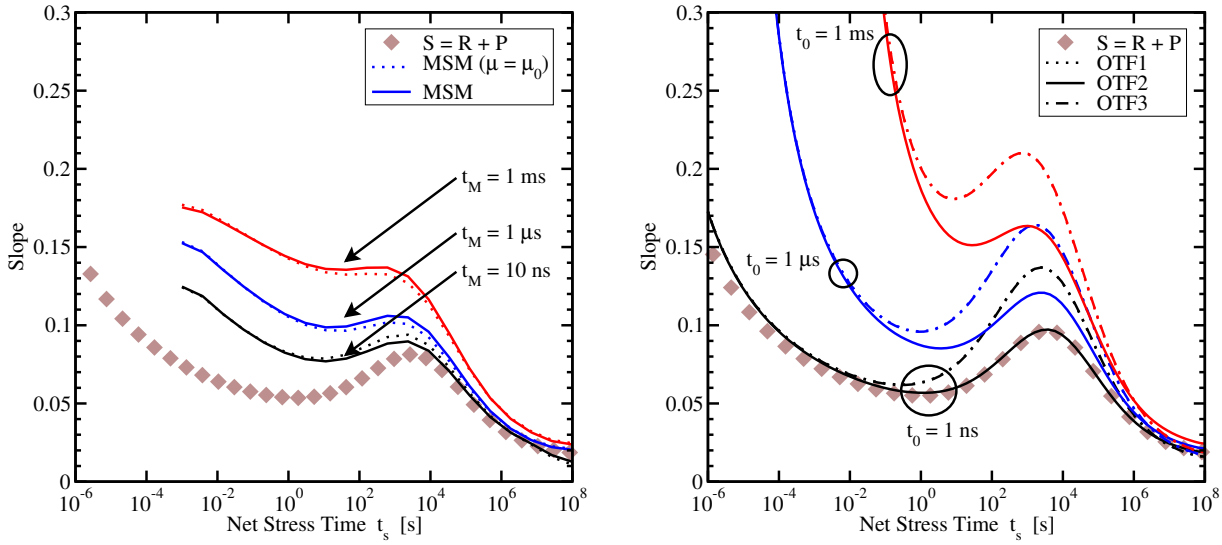


Figure 4.8: The observed power-law slope as a function of the stress time. **Left:** The MSM-method does not show severe mobility degradation errors and basically fits better the smaller the measurement delay t_M is. **Right:** Only an extremely fast OTF2 method with a hypothetical initial delay of 1 ns is able to reproduce the measurement data, while OTF1 and OTF3 are contaminated by t_0 , the mobility variation, and compact modeling errors ($\theta = \text{const}$, etc.).

while those including mobility variation yield a constant shift of +4 mV with respect to the real measurement. This is due to the fact that only interface states are assumed to affect mobility. As these states are considered permanent, only an upwards shift is obtained in the simulation. Moreover, the influence of the MSM-measurement delay, which strongly affects the extracted ΔV_{TH} , is very well described by the simulation results as shown in Fig. 4.7 (right).

A much more complex behavior is observed for the extracted degradation when using OTF methods, since OTF is seriously affected by the shift inherent in the first data point [40]. The larger the delay, the larger the distortion of the overall data gets, resulting in a problem similar to that caused by the MSM time delay, both depicted in Fig. 4.8. While OTF1 and OTF3 are prone to mobility changes, only OTF2 is uninfluenced by mobility changes. Only when furthermore presuming a t_0 of at least 1 ns small for the latter OTF2, a match of simulation and measurement is achieved.

4.4 Conclusion

An increasing number of publications is based on the assumption that there are two components responsible for NBTI: A fast, or universally recovering component on top of a slowly recovering or permanent component [6, 30, 49, 68]. However, the origin of the permanent and recoverable component has not yet been identified, as some authors, e.g. [6], claim that state interface states are permanent and oxide charges are recoverable, while others [31, 49] claim interface states to be solely responsible for NBTI. To reveal the responsible defects, two measurement techniques frequently used at present were studied in this chapter, the measurement-stress-measurement (MSM) routine and

the on-the-fly (OTF) method. Based on simulations augmented by suitable models for interface and oxide charges published in [39], it was tried to explain the experimental results of both techniques.

As expected, both techniques have their specific drawbacks. While the MSM-sequence suffers from an in-situ measurement delay, the OTF-techniques lack the initial reference measurement with the OTF1 and OTF3 extraction additionally being affected by mobility degradation. Moreover, the conversion routine to the threshold voltage shift introduces inaccuracies due to the simplifications made by the compact modeling, already explained in Chapter 2.3.

Nonetheless, the smaller systematic errors are found within the MSM routine. Despite its intrinsic delay, the time evolution of the recovery after BTI stress can be monitored most accurately. By using several MSM-sequences in a single measurement, the overall stress and relaxation behavior can be reconstructed as follows: Each recovery sequence can be fitted to the universal relaxation model by an optimization loop. The extracted permanent part P , i.e. the remaining degradation at the end of the extrapolated relaxation behavior, and the recoverable part R finally render the possibility to describe the influence of several stress parameters, like the temperature acceleration for BTI. While R seems to exhibit Arrhenius-like behavior with $E_A \approx 0.08$ eV independent of the stress time, P does not.

Finally, by using R and P it is also possible to explain the various values of extracted power-law stress exponents reported in literature. After a too long delay, i.e. $t_M > 1$ s, mostly P is left to monitor during the relaxation, while R has already disappeared. This makes the exponent depend on the delay time. Since higher delay times yield higher exponents, a lifetime extrapolation via such an exponent is questionable.

Chapter 5

Pulsed BTI Measurements

In the previous chapter various BTI stress tests were performed using the measurement-stress-measurement (MSM) and the on-the-fly (OTF) technique. Special attention was given to the fitting of the measurement data onto a universal relaxation law, yielding a separation of the degradation into a recoverable and a poorly recoverable or permanent component. Data gathered at different temperatures and stress voltages were found to follow a universal relaxation law. Interestingly both stress polarities, i.e. NBTI and PBTI stress on a pMOS, always resulted in a negative shift of the threshold voltage. Unfortunately, PBTI had been rarely discussed in literature until Liu *et al.* monitored a positive shift of the threshold voltage due to PBTI-stressed pMOS-devices [24], which contradicts the results presented by Grasser *et al.* [30].

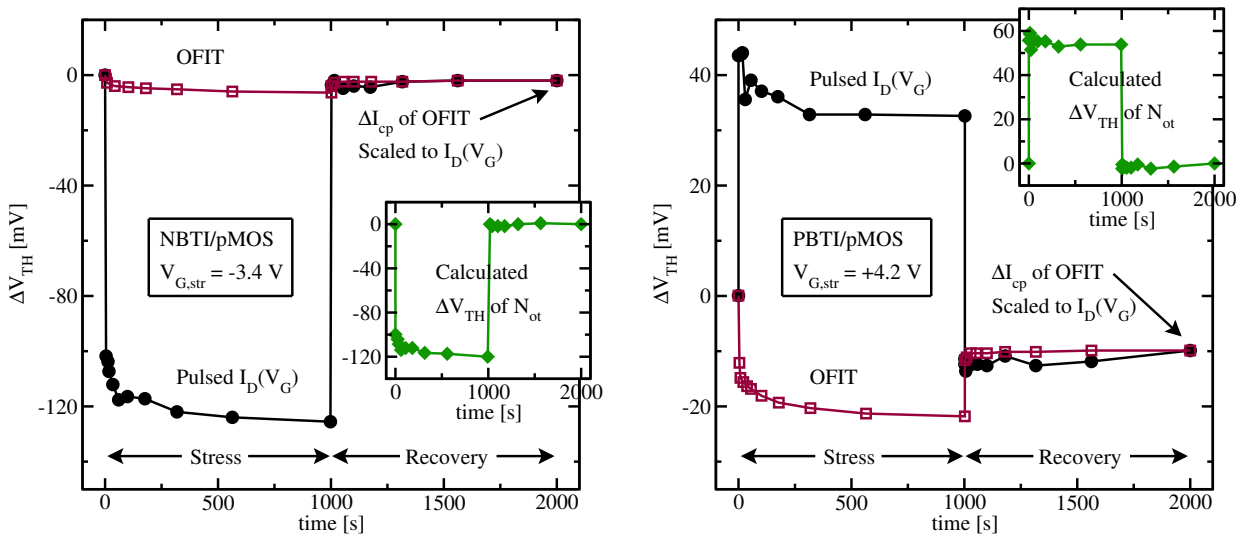


Figure 5.1: pMOSFETs monitored under 1000 s of NBTI (Left) and PBTI stress (Right) followed by 1000 s of relaxation. While ΔV_{TH} is measured by the FPM (open squares), ΔI_{cp} is measured by OFIT (solid circles). The fast pulsed $I_D(V_G)$ -characteristics reveal a negative shift of V_{TH} for NBTI (Left), while during PBTI a positive shift is visible (Right). At the end of the recovery phase the ΔI_{cp} curve is scaled to match the value of ΔV_{TH} . According to Liu *et al.* the difference between two curves (shown in the inset) yields the amount of contributing oxide traps. Data is taken from [24].

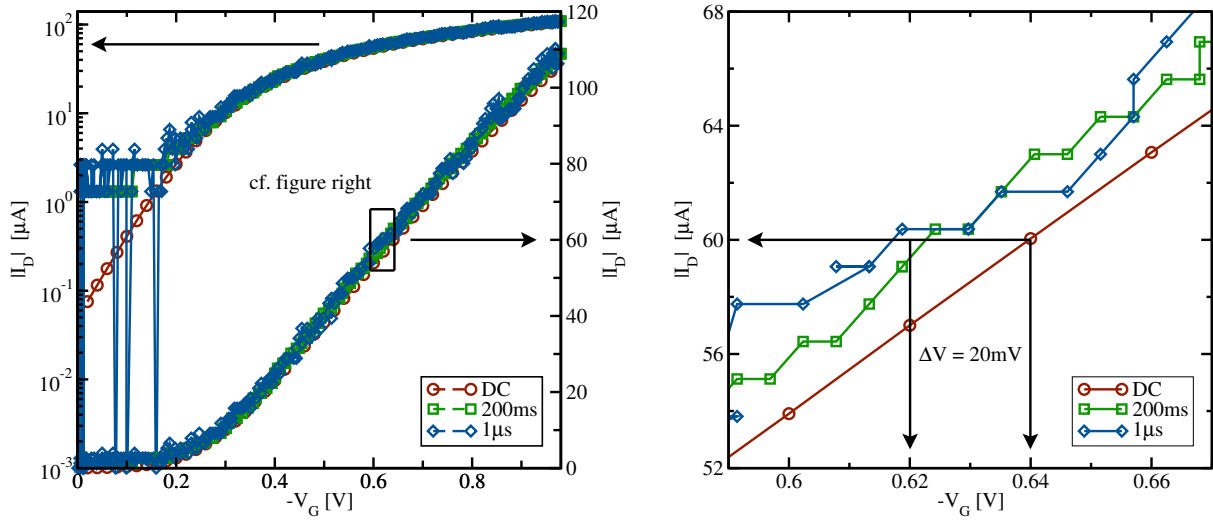


Figure 5.2: **Left:** Three different transfer characteristics. A DC-curve originating from a DSO with averaging acts as reference to $I_D(V_G)$ -characteristics obtained by two gate pulses with $t_P = 1 \mu\text{s}$ and $t_P = 200 \text{ms}$. Due to the limited resolution, especially the subthreshold region of the $I_D(V_G)$ is affected by quantization noise. **Right:** A close observation of the linear regime reveals different values of extracted V_{TH} . They differ by 20 mV from each other, as marked by ΔV which is on the order of the obtained degradation for 1000 s of PBTI stress, cf. Fig. 5.3 (left).

One reason of this discrepancy might be the fact that Grasser *et al.* used the OTF and the eMSM technique (cf. Chapter 2.3 and 2.1.3), while the two measurement techniques used in [24] are both based on the application of fast gate pulses: The newly developed on-the-fly fast charge pumping (OFIT) technique and the fast pulsed $I_D(V_G)$ -characteristics have been discussed in Chapter 2.5 and 2.2.1. The measurement results obtained by those two pulsed setups are only at a first glance interpreted in a correct way, as the ΔI_{cp} -curve obtained by OFIT is simply scaled to align the ΔV_{TH} -curve at the end of the recovery phase in [24]. Based on this alignment scheme depicted in Fig. 5.1, Liu *et al.* stated a fast oxide trap component (N_{ot}) corresponding to the difference of $\Delta V_{TH}(I_D(V_G)) - \Delta I_{cp}(\text{OFIT})$, which is shown in the inset of Fig. 5.1. Compared to that, the interface states are considered to recover only slowly. It was furthermore concluded that the fast oxide traps are responsible for the predominant part of V_{TH} -degradation in the fast pulsed $I_D(V_G)$ -characteristics only, since their influence during a DC measurement is drastically reduced due to the measurement delay. Consequently, this makes the interface states dominate the DC regime.

When taking a closer look at the pulsed $I_D(V_G)$ -characteristics of Fig. 5.1, a surprisingly huge offset of about 100 mV between the reference value and the first measurement point after 1 s of stress can be detected. As this already accounts for more than 75 % of the total degradation built up after 1000 s of stress, the high initial ΔV_{TH} seems to be at least questionable.

In order to determine to what extent interface states and oxide charges really contribute to the measurement signal, a more detailed study of the fast pulsed $I_D(V_G)$ and the OFIT technique, besides further measurements is needed. Especially the measurement delay of the setup in combination with its accuracy is of particular interest here.

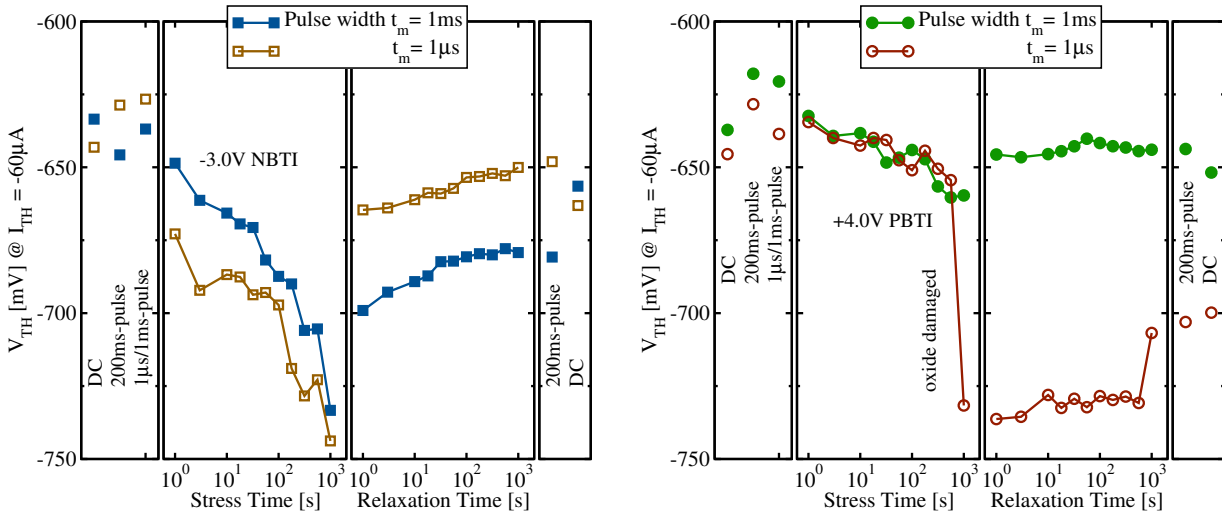


Figure 5.3: Fast pulsed $I_D(V_G)$ -measurements (FPM) performed on pMOS devices provided by IMEC after the method of Liu *et al.* Before FPM is applied using different pulse widths for NBTI/PBTI stress, V_{TH} is determined in three different ways, cf. Fig. 5.2. Both stress and recovery are interrupted 10 times within three decades ranging from 1 s to 1000 s for an FPM. Unfortunately, a high level of uncertainty is obtained by extracting the threshold voltage manually. **Left:** Applying NBTI stress yields sound results because of the higher signal-to-noise ratio and the expected negative shift of V_{TH} . **Right:** When performing PBTI stress again a negative shift of V_{TH} is found. This qualitatively supports the results of Grasser *et al.* presented in [30]. The suddenly appearing offset of -70 mV in-between the last two readout points during stress was assumed to be due to heavy oxide damage.

5.1 Pulsed $I_D(V_G)$ -Characteristics

Based on the large discrepancy between the initial reference and the very first measurement point visible in Fig. 5.1, different ways to extract a reference of V_{TH} are compared in Fig. 5.2. A DC-characteristic and a slow 200 ms-pulse sweep are both compared to the fast 1 μ s-pulse sweep which is used for the fast pulsed $I_D(V_G)$ -characteristics, cf. Fig. 5.4.

While the DC-curve is averaged and hence very smooth, the slow and fast pulses lack accuracy due to the missing averaging, as can be seen best in the subthreshold regime, which is very noisy. As depicted in Fig. 5.2 (right), setting the threshold current criteria to $I_{TH} = -60 \mu\text{A}$ (linear drain current regime), yields extracted values of V_{TH} differing in around 20 mV. This error is indicated as ΔV in Fig. 5.2.

The impact of the various transfer characteristics used to get an initial undegraded reference $V_{TH,0}$ to eventually measure ΔV_{TH} during stress is depicted for NBTI and PBTI stress in Fig. 5.3. Here the fast pulsed $I_D(V_G)$ -characterization using triangular 1 μ s- and 1 ms-pulses with zero pulse high-time after Li *et al.* (cf. Chapter 2.2.1) was applied to pMOS-devices with an $W/L = 10/0.35$ nm provided by IMEC¹.

For PBTI stress Fig. 5.3 (right) the determination of $V_{TH,0}$ delivers values which are of the same order of magnitude as the following degradation itself, cf. first and second subfigure. Depending on the chosen $V_{TH,0}$ -reference the determined degradation hence varies by a factor of two. The

¹As for extremely thin oxides (≤ 1.5 nm) direct tunneling occurs between the gate and the interface [52,93], 3 nm thick SiO_2 -dielectrics are used to avoid these tunneling currents.

same holds for the relaxation mode (third subfigure) and its V_{TH} -references taken at $t \approx t_{\text{rel}}$, the DC-characteristics and slow 200 ms-pulse (forth subfigure). Even more important is the fact that in contradiction to [24], the PBTI results do not exhibit a positive V_{TH} -shift at all, they solely show negative V_{TH} -shifts.

When the overall degradation becomes larger, as it is the case during NBTI stress (Fig. 5.3 (left)), the error induced by the reference decreases as expected. Unfortunately, the reason of the poor agreement of the differently extracted initial and post V_{TH} -values remains unclear. As these references do not indicate any systematic error, but seem to vary randomly, a different approach which is able to explain the deviating measurement results is needed.

5.2 Further Data Extraction Options

By focusing on the fast pulses only, the measurement setup is simplified. The next step is to examine the postprocessing of the measurement signal, as done in [24]. Thereby a very important fact becomes visible. The experimental output of the fast pulsed method, depicted in Fig. 5.3 (third and forth subfigure), was fitted by hand for each stress and relaxation time step.

To avoid the manual fitting, two methods to process all measurement data consistently are proposed. The first fits the data by using the SPICE level 1 compact model (2.7) from [37] and returns the threshold voltage as already introduced in Chapter 2.3. A second method combines the SPICE level 1 compact model with the constant current criterion of $I_{\text{TH}} = -60 \mu\text{A}$. Therefore the measurement data is first fit in the linear regime. Afterwards the extracted parameters β and θ are reinserted into (2.7) and reformulated as

$$V_{\text{G}} = \frac{V_{\text{D}}}{2} + \frac{I_{\text{D,lin}}}{\beta V_{\text{D}} - \theta I_{\text{D,lin}}} + V_{\theta}, \quad (5.1)$$

with $I_{\text{D,lin}} = I_{\text{TH}}$. These two extraction schemes will be referred to as “avg” and “avg + I_{TH} ” in the following. Before addressing their differences further, another point needs to be discussed.

The pulse polarity differs for NBTI and PBTI during the stress phase. This becomes obvious in Fig. 5.4, where the standard FPM scheme is examined. The pulse slope used for the V_{TH} -extraction is highlighted with circles for the three modes of operation: the initial, the stress, and the relaxation phase.

The transistor is usually driven from accumulation towards inversion and back, i.e. a falling pulse edge is followed by a rising pulse edge. Only during NBTI stress the polarity of this pulse is reversed. What at a first glance seems to be irrelevant, namely which edge is chosen for the extraction, actually turns out to be significant. The $I_{\text{D}}(V_{\text{G}})$ -curves extracted from the two pulse edges forming each pulse do not necessarily coincide as primarily assumed, rather they show a hysteresis. This hysteresis originates from slightly stressing or relaxing the device through the pulsed measurement itself. Consequently the pulse hysteresis influences the extracted values of the threshold voltage.

In the following all meaningful pulse edge combinations for both NBTI and PBTI are schematically compared in Tab. 5.1. The extracted values of V_{TH} are displayed in Fig. 5.5 (left) for a 1 ms-pulsed NBTI-FPM. Depending on which edges are used, the degradation is either overestimated (rising or first pulse edge) or underestimated (falling or second pulse edge). The first case is due to the fact that the device obviously still relaxes while V_{TH} is determined, while the latter already suffers from too long delay times, i.e. the time of one pulse edge (1 ms) is already missed.

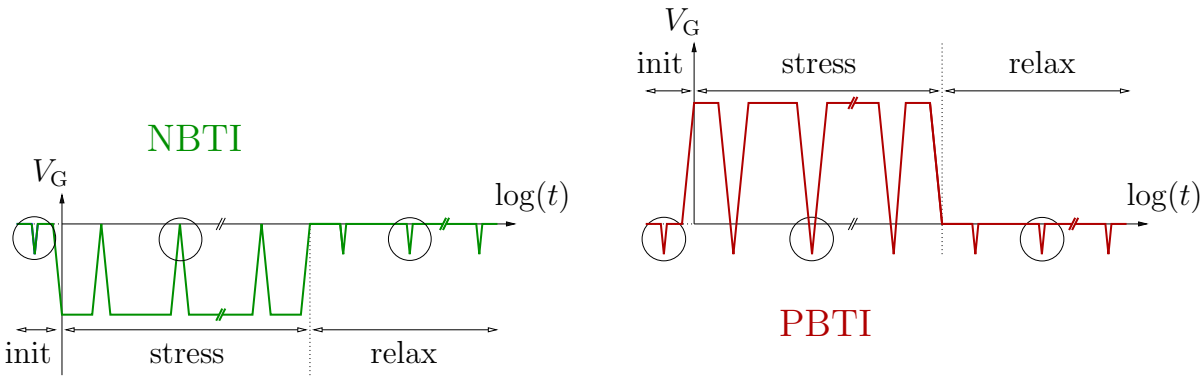


Figure 5.4: FPM performed with triangular gate pulses. The circles mark the different pulse shapes during the initial phase, the stress phase, and the relaxation phase. **Left:** NBTI stress relaxation sequence. **Right:** PBTI stress relaxation sequence.

NBTI	init	stress	relax	PBTI	init	stress	relax
Rising pulses	✓	✓	✓	Rising pulses \equiv Second	✓	✓	✓
Falling pulses	✓	✓	✓	Falling pulses \equiv First	✓	✓	✓
First	✓	✓	✓	Permutation #1	✓	✓	✓
Second	✓	✓	✓	Permutation #2	✓	✓	✓
Both averaged	✓	✓	✓	Both averaged	✓	✓	✓

Table 5.1: Each pulse of the FPM can be split into two pulse edges, a rising and a falling part. When describing the three different phases of the initial reference, the stress, and the relaxation measurement, which are schematically depicted in the figure above, the highlighted permutations are feasible.

During relaxation the same pulse form is used for both NBTI and PBTI. Therefore the hysteresis does not affect the extracted results. As expected, smoother results are obtained by averaging the pulses.

5.2.1 Determination of the Fitting Region

It is a challenge to determine the range of measurement data within the FPM pulse, which is further used to extract V_{TH} . This is because on the one hand side the noisy subthreshold region should be avoided, but at the same time a preferably large range of data points is required for the fitting algorithm. This balancing act is illustrated in Fig. 5.5 (right). A good compromise is found by using all data points between $V_G = -0.32$ V, where the signal-to-noise ratio is still reasonable high, and the stress level of NBTI at $V_G = -3.0$ V. This range provides the best possible agreement of the manual fitting and the two proposed extraction schemes, avg and $avg + I_{TH}$.

5.2.2 Impact of the Pulse Amplitude

Unfortunately, the different FPM-setups for NBTI and PBTI featuring different pulse amplitudes do not give comparable qualitative results for degradation and relaxation (cf. Fig. 5.4). During an FPM-pulse, always the same number of data points per 1 V is recorded to guarantee a constant resolution of pulse amplitude per time, i.e. a constant slope. However, V_{TH} -extraction using (5.1) strongly depends on the data range of the gate pulse. The usable range was already determined to

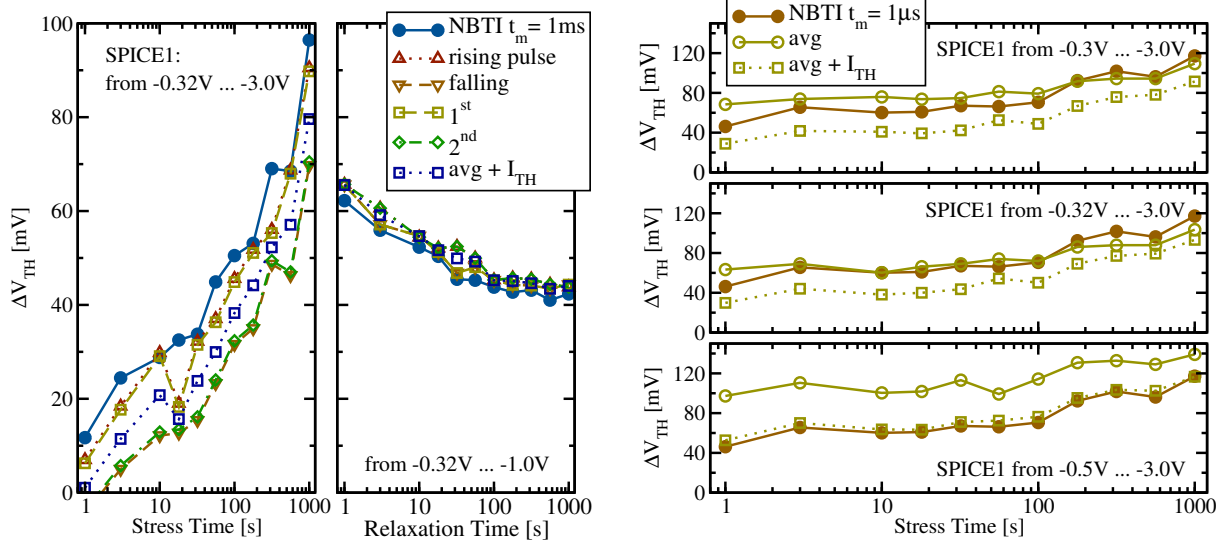


Figure 5.5: Changes of the threshold voltage of the SPICE level 1 model fit to FPM after [24,51]. **Left:** The results of the manual extraction routine, which is applied according to [24,25,51] (filled symbols), are compared to the proposed “avg + I_{TH} ” extraction scheme. By using various pulse edges of the measurement data as displayed in Tab. 5.1, more or less smooth ΔV_{TH} -curves are obtained. The extraction which uses both pulse edges (averaged pulse) yields the best possible results that can be obtained for FPM. Due to the hysteresis between the rising and falling edge all other pulse combinations give barely acceptable results. **Right:** The V_G -range used to extract V_{TH} properly is restricted by the noise in the subthreshold regime on the one hand and by the maximally accessible inversion regime on the other hand. The optimum V_G -range goes from -0.32 V to -3.0 V.

be less than 3 V wide for NBTI stress, but is unavoidably even more limited for PBTI stress. That is because it is important to avoid any NBTI stress during a PBTI-FPM, since the interplay between NBTI and PBTI is not yet understood and therefore not distinguishable at the present day. In fact, the readout pulse during PBTI mainly covers the accumulation regime and just records the onset of inversion; it is not allowed further towards more negative bias. This yields a very limited range of -0.32 V down to -1.0 V for the V_{TH} -extraction during PBTI.

Unfortunately, the noisy measurement data in combination with the avg fitting algorithm results in an oscillating stress curve, depicted clearly in Fig. 5.6 (left). Applying the avg + I_{TH} -extraction heavily reduces this oscillation as the current criterion in the linear regime ($I_D = 60 \mu\text{A}$) is less sensitive to a change of the slope of the $I_D(V_G)$ -characteristics. For NBTI, the strongly differing degradation values after 1 s as well as the differing slopes of the degradation curves for the manual fitting, the avg-, and the avg + I_{TH} -extraction can be explained by the different mobility degradation of the corresponding V_{TH} . As the extraction scheme of V_{TH} is at least comparable for the manual and the avg + I_{TH} -extraction, because both are extracted in the linear drain current regime, the corresponding degradation and relaxation is also similar.

5.2.3 Varying Pulse Rise/Fall Times

An important point has not yet been discussed so far. When investigating short pulsed BTI measurements such as FPM, the time-resolution of the measurement equipment is of utmost importance.

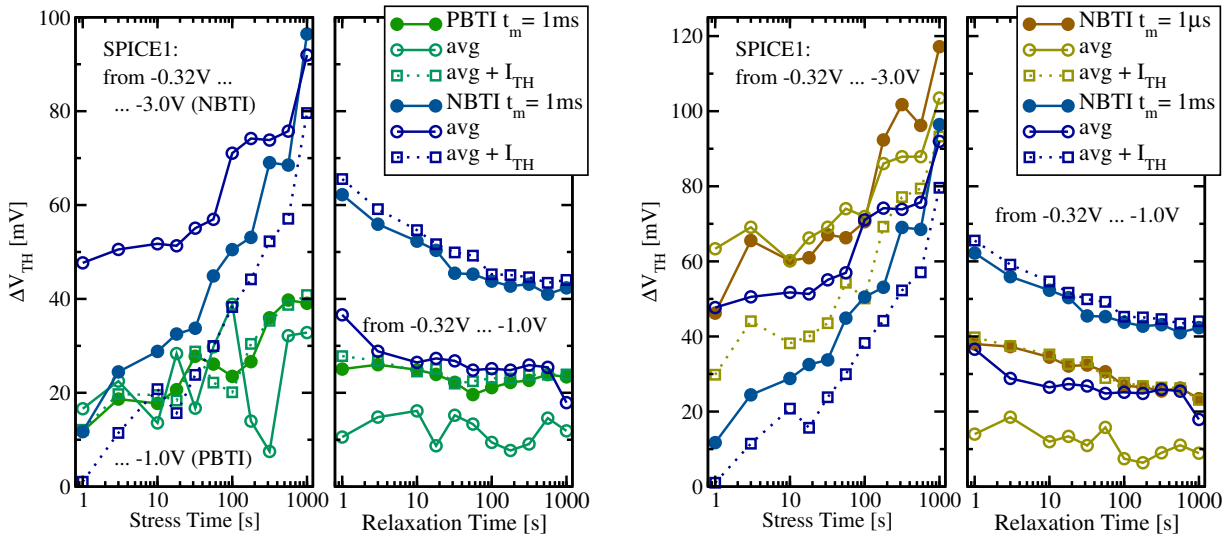


Figure 5.6: Fast pulsed $I_D(V_G)$ -measurements after [24, 51]. **Left:** In order to prevent any effects caused by NBTI stress, the possible pulse amplitude during PBTI stress is limited. Hence fitting these pulses yields very noisy output in contrast to NBTI, whose pulses are not strictly constrained. These different limits of the usable pulse amplitude make the FPM routine less applicable to PBTI stress compared to NBTI stress. **Right:** The comparison between longer and shorter pulse times reveals no surprise, as the longer 1 ms-NBTI-pulse gives the best match to the manually extracted one.

To determine the limits of the used setup, a short 1 ms- and a very short 1 μ s-pulse mode² are compared for NBTI in Fig. 5.6 (right). Not surprising, the extracted values using the 1 ms-pulses are smoother than those using the 1 μ s-pulses. This is due to the best signal-to-noise ratio of all four performed FPM (NBTI/PBTI using 1 μ s/1 ms), previously depicted in Fig. 5.3.

5.2.4 Consequences

The two described extraction schemes of the FPM-method, namely the avg- and the avg + I_{TH} -extraction, have shown that fully automated handling of a dataset helps to consistently compare experimental results. However, the performed measurements also underlined the fact that even with proper fitting/smoothing methods to avoid noise as much as possible, the practicability of the measurement routine has to be checked first, especially when dealing with different pulse polarities for NBTI and PBTI.

In the case of PBTI a detailed characterization via FPM is simply not possible because the pulse settings are not suitable for both PBTI stress and recovery characterization in a single measurement. The settings are usually a compromise between maximizing the data range for the $I_D(V_G)$ -characteristics on the one hand and preventing the device from undesired NBTI stress on the other hand. The latter case occurs when the device is driven too far into inversion. Despite these drawbacks the trend of the degradation can be determined, cf. Fig. 5.3. It features a negative shift of the threshold voltage, comparable to NBTI but smaller. So far this refutes the existence of electron tunneling as stated in [24], but unfortunately the actual type of defects contributing to BTI still remains unclear. Therefore the measurement technique proposed at the beginning of

²The pulse time here corresponds to the added rise and fall time of the pulse.

this chapter using charge pumping will be investigated next. Special emphasis is again put on the measurement method itself.

5.3 Experimental Identification of Defects

It was already shown in Chapter 2 that the experimental access to BTI is extremely challenging due to the rapid recovery of the degradation, setting in as soon as the stress is removed. In particular, it has been observed that when after NBTI stress the device is positively biased, a considerable part of the recoverable component is lost [6, 62, 77, 94]. Until recently, this has been explained by the detrapping of holes [6, 62], while interface states have been assumed to only change their occupancy but do not recover. Unfortunately, this makes any experimental assessment of the defects contributing to BTI very complex.

A quite striking result obtained with on-the-fly charge pumping (OFIT) measurements is that in contradiction to charge pumping (CP) measurements, OFIT data suggest a considerable amount of fast initial recovery of interface states. It has to be noted that this fast initial recovery is not explicitly measured, but is only inferred from the differences between the last stress and the first recovery measurement. The correctness of this assumption heavily affects the understanding of the short-term behavior of interface states. It is hence a major topic to clarify this subject as it is the prime requisite for the development of a reliable model.

By performing CP and OFIT measurements on different technologies as described in Chapters 2.4 and 2.5, the issue whether interface states do recover quickly (<1 s) or not is resolved in the following.

5.4 OFIT versus CP

As in conventional CP measurements, care has to be taken that parasitic tunneling currents and geometry effects do not pollute the measured charge pumping current I_{cp} . The first problem is even more severe in the OFIT technique since there the low level gate voltage equals the stress voltage, resulting in excessive tunneling in thin oxides [51]. In order to avoid these problems, also large-area devices with thick oxides (30 nm) are used. As shown in Fig. 5.7, the measured I_{cp} during stress and recovery are qualitatively identical for three completely different technologies (30 nm thick SiO₂, 3.5 nm thin SiO₂, and SiON).

Quite remarkably, continuous application of OFIT pulses (as well as CP measurements) has a dramatic impact on both the stress and the recovery characteristics. In particular, with 3 measurements per decade of relaxation time, I_{cp} is quasi constant during recovery, while up to 100 measurements per decade in time result in approximately 80% recovery of I_{cp} .

Another fact is that the first OFIT measurement point during stress is already responsible for at least 30% of the total degradation. Likewise, the first measurement taken during recovery at 300 ms already shows 30% recovery while the rest of the recovery depends basically on the number of measurements per decade. Though not shown here, the same behavior is obtained for CP. To be able to understand how recovery works here, a deeper analysis of the charge pumping technique is needed.

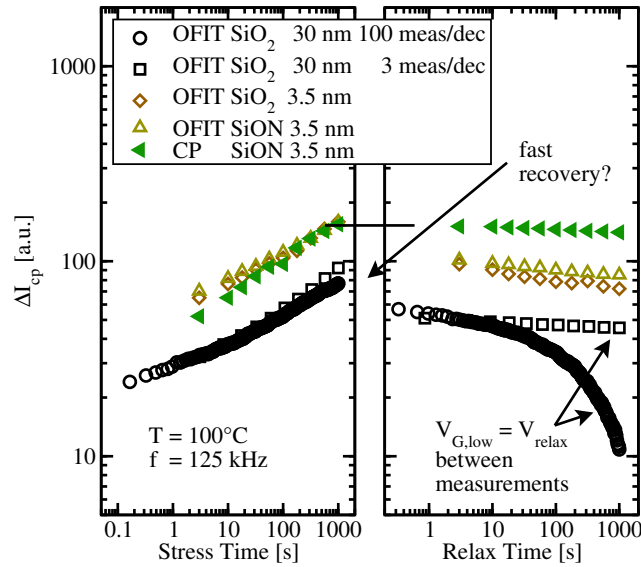


Figure 5.7: Comparison of OFIT and CP results. For OFIT the offset between end of stress and beginning of relaxation is comparable for different technologies and geometries (not shown). Previously this was explained by fast recovery. This fast recovery is absent in CP (indicated for SiON). Furthermore, continuous gate pulsing affects both stress and relaxation, causing a faster recovery of interface states with increasing number of measurements (black circles vs. black squares, both of 30 nm OFIT).

5.5 Analysis of the OFIT Technique

As described in [95, 96], a constant base-level CP measurement with $V_{\text{Base}} = 2\text{ V}$ is performed using a gradually increasing pulse amplitude ΔV_G . Until the desired stress level is reached, starting from -1 V down to -18 V , the pulse slopes have to be kept constant to obtain comparable results. Constant pulse slopes ensure that the upper and lower energy boundaries of the active energy interval remain unchanged when ΔV_G increases [50]. Due to a constant pulse slope the amplitude of ΔV_G is proportional to the pulse rising (often referred to as leading) and falling (trailing) time. Given the additional requirement of a constant duty cycle, the rise and fall times have to be adapted at every voltage step within the CP measurement to obtain the proper charge pumping current I_{CP} . Since it is inevitable to change both the pulse width and also the rise and fall times one has to ask for the potential pitfalls: Are OFIT-data obtained during stress and relaxation comparable? If that is not the case, is there some possibility to correct this nonconformity? These questions will be examined in the following.

Starting with Fig. 5.8 (left) the two large arrows pointing up and down reveal some important aspects of the temporal evolution of the pulses during a CP measurement. The charge pumping current I_{CP} at stress conditions ($V_{\text{G,low}} < -3\text{ V}$) differs a lot when compared to that obtained during relaxation ($V_{\text{G,low}} = -1\text{ V}$). The higher the NBTI stress conditions, the larger the I_{CP} -signal becomes. This can be partly attributed to the desired effect of using the measurement setup to also stress the device. However, it cannot fully account for the observed behavior.

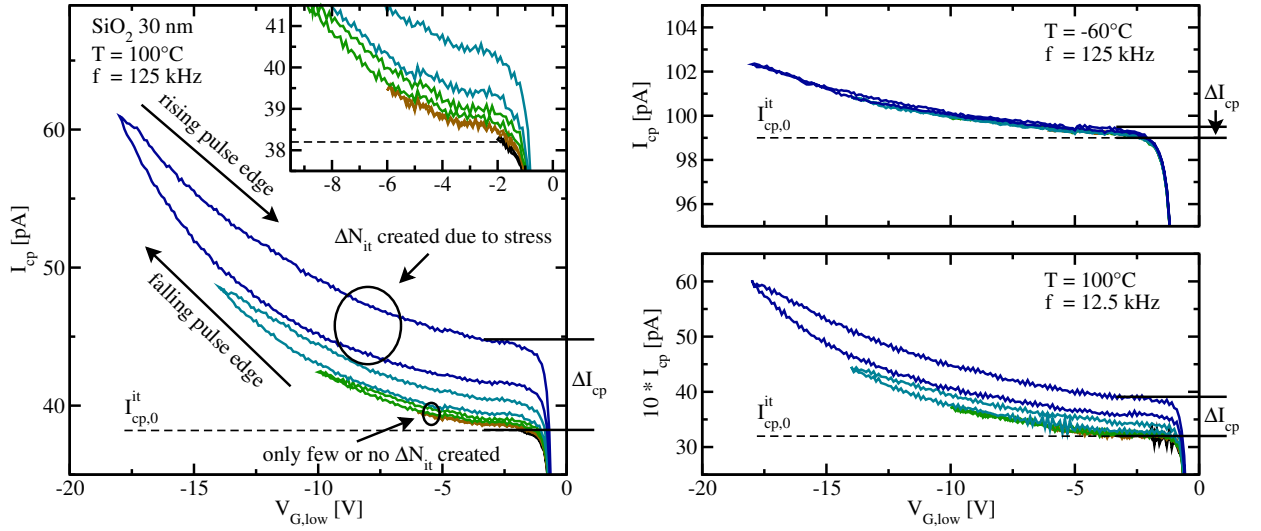


Figure 5.8: **Left:** Charge pumping current I_{cp} for different pulse amplitudes as observed in constant base-level CP measurements with $V_{Base} = 2\text{ V}$ and a gradually increasing pulse amplitude $\Delta V_G = V_{Base} - V_{G,low}$ from $V_{G,low} = -1\text{ V}$ down to $V_{G,low} = -17\text{ V}$. I_{cp} shows a significant hysteresis. If I_{cp} is evaluated at the falling pulse edge, the lower branch of the curve is traversed. Evaluation of the rising pulse edges gives the upper branch. However, the contribution of slow oxide states and an additional hysteresis (marked with ΔI_{cp}) are clearly visible for increasing pulse amplitudes. This implies that depending on the pulse amplitude, I_{cp} will contain contributions of both, interface and oxide states. Provided only interface states are available, I_{cp} should be independent of the pulse amplitude (dashed line of $I_{cp,0}^{it}$). **Top Right:** At low temperatures the hysteresis is negligible (less than 1%) and the contribution of slow oxide traps is reduced. **Bottom Right:** At low frequencies and high temperatures the contribution of oxide traps increases due to the increased rise and fall times. A comparable if not equal part of interface states constituting ΔI_{cp} can be identified for different frequencies but equal temperature when checking against the left figure. Following these results at least part of the defects must vary with temperature or frequency. For better comparability, the data at 12.5 kHz are scaled to the reference frequency ($f_{ref} = 125\text{ kHz}$).

5.5.1 Dependence on Gate Voltage Low-Level

To compare stress and relaxation properly, the initial charge pumping signal $I_{cp,0}^{it}$ should stay constant over the whole considered low level gate voltage V_G -region. Hence, under the assumption that only interface states contribute, I_{cp} should actually become independent of $V_{G,low}$ as soon as the strong inversion regime is reached. This $I_{cp,0}^{it}$ is marked by the dashed line in the left of Fig. 5.8. However, as demonstrated previously [43,44], I_{cp} continues to increase, albeit at a much slower rate. This increase with ΔV_G is routinely attributed to slower oxide traps ΔN_{ot} and $I_{cp} = I_{cp}^{it} + I_{cp}^{ot}$ [47,97]. So, regardless of the amount of degradation, I_{cp} varies as function of $V_{G,low}$. This fact has to be taken into account for a meaningful comparison of stress and relaxation CP data.

5.5.2 Hysteresis due to Stress

When $V_{G,low}$ is lowered towards the stress voltage, as required in the OFIT technique, I_{cp} extracted from the rising and falling pulse edges start to deviate, introducing a hysteresis. The hysteresis is only visible for larger pulse amplitudes, indicating degradation (marked with ΔN_{it} and ΔI_{cp})

due to stress. While the impact of the oxide traps visible during medium $V_{G,\text{low}}$ appears to be fully recoverable, the component causing the hysteresis is not. This can be seen in Fig. 5.8, where I_{cp} increases during subsequent measurements performed on the same device. We attribute this hysteresis to the creation of additional interface states due to NBTI stress at $V_{G,\text{low}} = V_{\text{str}}$ [78]. Starting at -2 V there is nearly no stress. The deeper the device is stressed into inversion the larger the hysteresis becomes, resulting in an increased offset for the next pulse. The total hysteresis at a certain stress level hence not only consists of the hysteresis of the momentary charge pumping measurement but depends on the previous measurements³.

As displayed in the inset in Fig. 5.8 (left), the very first pulses are almost free of stress (no hysteresis, $\Delta I_{\text{cp}} = 0$) and hence the deviation of I_{cp} from $I_{\text{cp},0}^{\text{it}}$ is entirely due to oxide traps. Only a negligible amount of interface states ΔN_{it} are created by the measurement process. The hysteresis-free area will be discussed in more detail in the next section.

When the experiment is repeated at a lower frequency (see bottom right of Fig. 5.8), one finds that the interface state contribution can be scaled to the reference frequency ($f_{\text{ref}} = 125\text{ kHz}$) [44]. This is compatible with the fact that the stress duration is practically independent of frequency. On the other hand, the recoverable oxide trap contribution to I_{cp} depends on frequency, consistent with the idea that the lower the frequency (corresponding to more time per pulse) the more oxide traps can contribute to I_{cp} [95].

Finally, at a low temperature, displayed at the top right of Fig. 5.8, practically no hysteresis is introduced (no NBTI stress) and also the oxide trap contribution is reduced, consistent with the idea that these traps are due to a thermally activated tunneling mechanism [98] rather than elastic (and thus temperature-independent) hole tunneling [94].

5.6 Extrapolation of Oxide Trap Contribution

As demonstrated above, during an OFIT measurement a distortion of I_{cp} due to oxide charges and due to the creation of defects during the low-level is monitored. In order to analyze this distortion, $\hat{V}_{G,\text{low}}$ is determined to be the lowest value of $V_{G,\text{low}}$ at which no hysteresis is observed. The dataset $V_{G,\text{low}} > \hat{V}_{G,\text{low}}$ is then used to extrapolate the impact of oxide charges ΔN_{ot} down to the stress-level. It is not possible to obtain this information from the stress pulse because of the contribution of both parts ΔN_{it} and ΔN_{ot} . Quite remarkably, the data [78] can be fit by a quadratic polynomial, consistent with our NBTI experiments where we also observe a quadratic (E_{ox}^2) dependence of the hole-trapping component [18, 98, 99]. The hole-trapping theory developed in [98] was applied to our data and excellent agreement was obtained. The difference between the actual signal ($I_{\text{cp}}^{\text{it}} + I_{\text{cp}}^{\text{ot}}$) and the extrapolated curve in Fig. 5.9 finally gives ΔN_{it} .

In Fig. 5.9 the extraction algorithm for ΔN_{ot} and ΔN_{it} is demonstrated. Stress and relaxation pulse responses both consist of two branches, one falling and one rising, as marked by arrows. In the falling branch, $V_{G,\text{low}}$ varies from 0 V to -17 V . In the rising branch, $V_{G,\text{low}}$ varies from -17 V to 0 V . Only pulses with constant $I_{\text{cp}}^{\text{rise}} - I_{\text{cp}}^{\text{fall}}$ (or even without a hysteresis, i.e. $I_{\text{cp}}^{\text{rise}} - I_{\text{cp}}^{\text{fall}} = 0$) can be used to create an extrapolation guess for higher $V_{G,\text{low}}$. This ‘safe window’ ranges from 0 V to -8 V , where both branches are indistinguishable.

The extracted components for different temperatures and frequencies are given in Fig. 5.10. The additionally created oxide traps ΔN_{ot} depend on frequency as well as on temperature and clearly

³Note that all pulses in the graph are performed on the same device to ensure comparability.

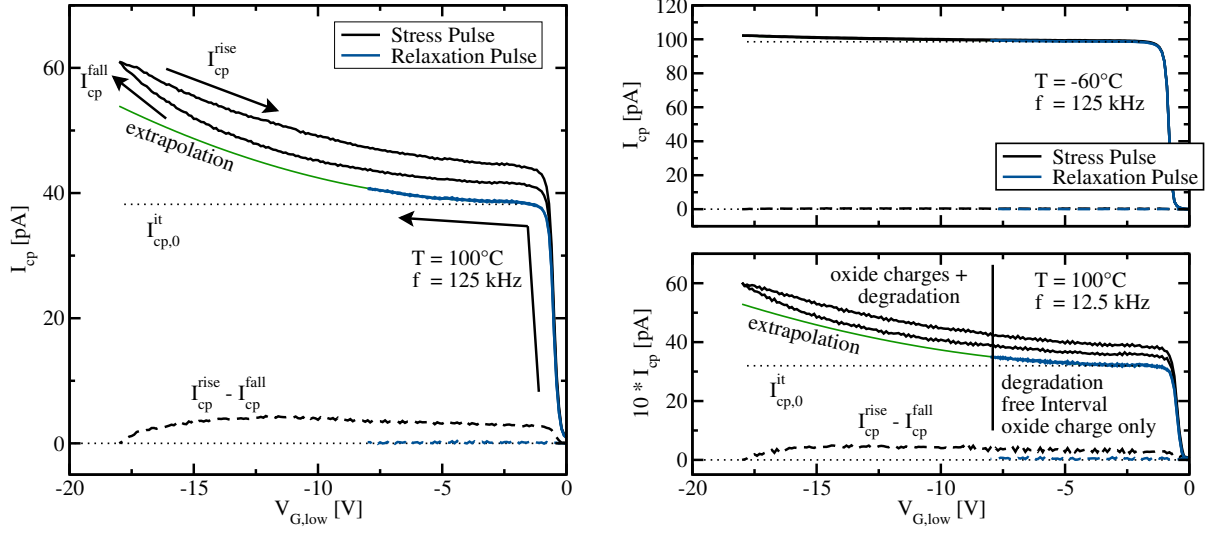


Figure 5.9: Left: Charge pumping current I_{cp} for the stress pulse ($V_{stress} = -17\text{ V}$) and the relaxation pulse ($V_{relax} = -8\text{ V}$), shown in Fig. 5.8 (left). To decompose the contribution of oxide charges and additional interface states we look at the difference $I_{cp}^{rise} - I_{cp}^{fall}$. In the range $-8\text{ V} < V_{G,low} < 0\text{ V}$, this difference is constant, implying no additional creation of interface states. From this ‘safe window’ we extrapolate to the minimum low-level to estimate the contribution due to oxide charges. Note that the first branches I_{cp}^{fall} of the stress and relaxation pulse differ from each other due to pre-stress pulses between $V_{G,low} = -8\text{ V}$ and $V_{G,low} = -17\text{ V}$. In fact, when using fresh devices for each measurement all I_{cp}^{fall} would coincide. **Top Right:** Lower temperatures simplify the extrapolation due to the absence of degradation. Here the full range of pulse amplitudes can be used to verify the extrapolation down to deep inversion. The missing hysteresis indicates the absence of additional oxide states in deep inversion at low temperatures. **Bottom Right:** Noise complicates this procedure at low frequencies. Data are scaled to $f_{ref} = 125\text{ kHz}$.

show $V_{G,low}^2 \sim E_{ox}^2$ behavior. The hysteresis due to additionally created traps ΔN_{it} is independent of frequency, but strongly dependent on temperature.

5.7 Simulation of the Charge Pumping Current

To approximately account for the above mentioned temperature and field activated tunneling process, a modified Shockley-Read-Hall (SRH) model⁴ is used within our device simulator Minimos-NT [89]. The SRH-capture-rates are multiplied by

$$\exp\left(\frac{E_{ox}^2}{E_{ox,ref}^2}\right) \exp\left(-\frac{\Delta E_B}{k_B T}\right) \quad (5.2)$$

where E_{ox} is the electric field in the oxide, $E_{ox,ref}$ is a reference value, ΔE_B the multi-phonon emission barrier and $k_B T$ the thermal energy. ΔE_B can be characterized by a Gaussian distribution with the mean energy $\Delta E_{B,mean}$. When setting the parameters some points need to be considered in order to end up with a physically appropriate model:

⁴Additive rates of the generation and recombination model are introduced in order to model the oxide traps.

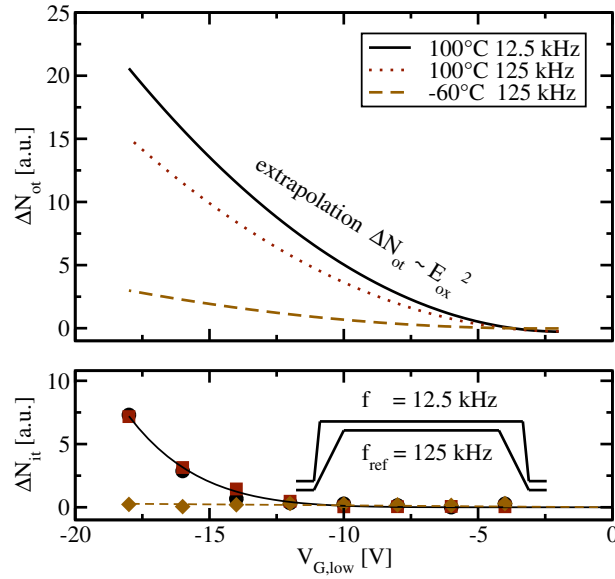


Figure 5.10: The extracted oxide state density (**Top**) and additional interface state density (**Bottom**). The change of oxide trap density ΔN_{ot} follows E_{ox}^2 , and depends on the frequency as well as on the temperature. The hystereses displayed in the previous figures are due to additionally created traps, ΔN_{it} , which are independent of the frequency but strongly dependent on the temperature.

1. The first exponential factor in (5.2) models the bias dependence. It is very sensitive to changes of $E_{ox,ref}$ due to the squared exponent, leading to a very small range of valid $E_{ox,ref}$ values. This $E_{ox,ref}$ reference field acts as a scaling factor.
2. When setting the barrier too low, the oxide traps contribute to the interface trap signal as the second factor approaches unity. Setting the barrier too high leads to very low rates, effectively eliminating the contribution of oxide traps.
3. The distribution of ΔE_B determines the dependence of I_{cp} on $V_{G,low}$. Increasing the mean of the distribution at ΔE_B increases the mean capture/emission-time constants. Since with constant-slope pulses higher pulse amplitudes ΔV_G require longer pulse durations, increasing the mean ΔE_B shifts the point from which a significant contribution of oxide traps ΔN_{ot} can be observed to higher pulse amplitudes. On the other hand, broadening the distribution of ΔE_B (increasing the variance) also broadens the distribution of time constants, observable as broadening the range of $V_{G,low}$ where I_{cp} increases.
4. Lastly, the distribution determines how strong some oxide traps contribute to the I_{cp} -signal in each time-interval of the pulse. To achieve a smooth quadratic behavior as observed in the experiments, Fig. 5.11, a broad Gaussian peak over a wide range of energies is required ($\Delta E_{B,mean} = 1 \text{ eV}$, $\Delta E_{B,\sigma} = 0.5 \text{ eV}$), consistent with other NBTI experiments [98, 100].

The final simulation results are depicted in Fig. 5.11. As the simulation treats the CP measurement process as stress-free, no additional interface traps are created and only the oxide-charge part is visible. With the thermally activated barrier the increasing I_{cp} can be described.

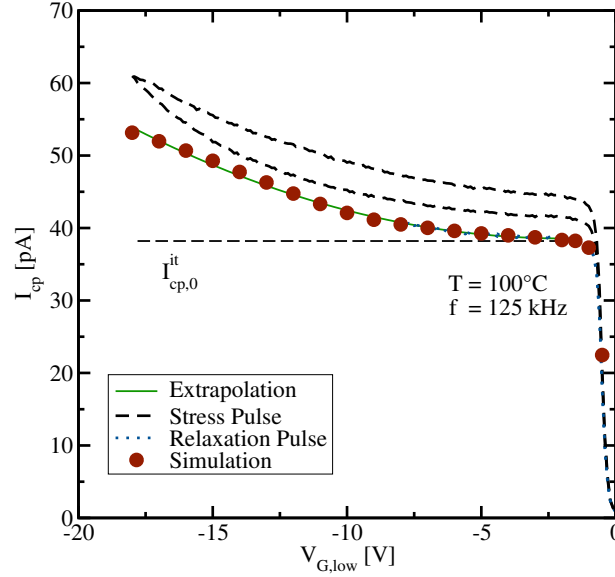


Figure 5.11: The contribution due to oxide traps can be well described using the model suggested in [98]. This model assumes that hole-trapping is possible via a multi-phonon process implying a thermally activated barrier $\exp(-\Delta E_B/k_B T)$ and an $\exp(E_{ox}^2/E_{ox,ref}^2)$ field dependence. It is stipulated that the simulation only describes oxide traps and interface traps without applied stress conditions. Additional interface traps due to NBTI stress are missing in the simulation because of a constant number of interface traps for each simulation point (solid circles).

5.8 Results

Based on the previous results we are now able to better understand the charge pumping current I_{cp} measured with the OFIT sequence. The presence of additional charges contributes to the signal when the pulse amplitude ΔV_G is increased. A large spread of time constants larger than that of the interface states is necessary to explain the results. By assuming oxide traps with a distributed thermally activated barrier one is able to explain the measurement results with good accuracy. Whereas interface states seem to not respond to an increasing electric field and due to their small time constants account for I_{cp} at low (10 kHz) and high frequencies (1 MHz), the oxide traps are by far slower due to the assumed barrier ΔE_B they have to surmount. That is why oxide traps only affect I_{cp} at lower frequencies, i.e. 10 kHz.

The particularly troublesome part is the application of the OFIT technique during the stress phase, where both oxide traps ΔN_{ot} and additionally created interface states ΔN_{it} add to I_{cp} . These contributions are absent during the initial reference measurements and during the OFIT recovery measurements both taken at $V_{G,low} = V_{rel}$. This has fundamental consequences on OFIT measurements: Initially, a reference I_{cp} is recorded. Following this reference measurement, the gate voltage low-level $V_{G,low}$ is switched to V_{str} . Due to the much larger ΔV_G now a significant contribution of I_{cp}^{ox} is obtained. Furthermore, with the large pulse amplitude, additional interface states are created, which is the intended effect of this OFIT measurement. However, without this the additional increase in I_{cp} due to oxide traps must not be attributed to interface states created by degradation. Consequently, I_{cp}^{ot} needs to be corrected in the measurement data. Using the mentioned extrapolation method of $\Delta N_{ot} = A E_{ox}^2$ reveals that the 30% initial increase in I_{cp} is entirely due to

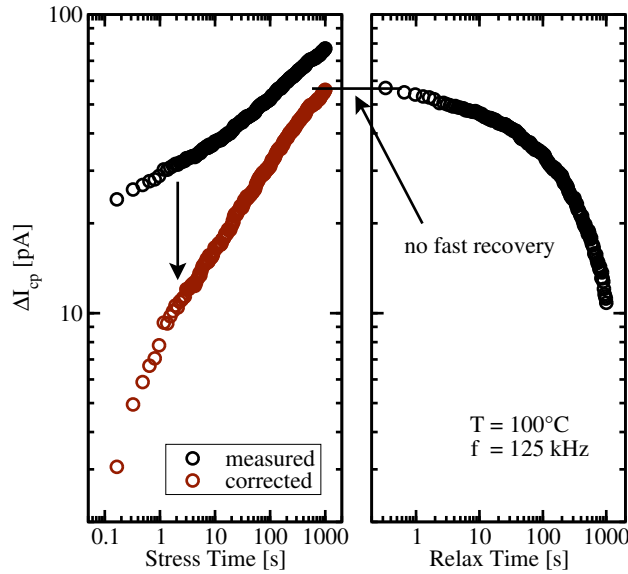


Figure 5.12: Oxide traps lead to a spurious increase in the charge pumping signal during stress. Using the scheme developed for Fig. 2.7, a corrected I_{cp} is obtained. The smooth transition between the corrected I_{cp} during stress and the I_{cp} during recovery suggests that no fast recovery takes place.

oxide traps. The corrected last stress value in Fig. 5.12 is identical to the first value at the recovery, leading to the conclusion that no fast interface state recovery occurs.

5.9 Conclusion

With the help of the pulsed measurement techniques presented in this chapter, the FPM and the OFIT technique, it was tried to reveal the recovery accounting for BTI. The FPM technique provides a sophisticated way to monitor the degradation of a MOSFET when the data is extracted consistently. Nevertheless, it suffers from a serious shortcoming: It is not suited for PBTI, since the determination of V_{TH} requires to reach deep inversion which should be avoided as far as possible in order to avoid a superposition of NBTI and PBTI. This renders a precise and equivalent comparison of NBTI and PBTI impossible. In spite of that the previously observed positive V_{TH} -shift during FPM which was explained by electron tunneling could not be reproduced. Quite the contrary, solely negative V_{TH} -shift was observed for both NBTI and PBTI.

The second proposed pulse technique performs constant base-level charge pumping measurements to access the dynamics of interface states. Thereby it was found that the charge pumping current I_{cp} is not constant in the inversion regime, but increases due to slow oxide traps. As a consequence, the data gathered during stress and recovery phases for the OFIT measurement technique is fundamentally different and must not be directly compared, as for example done in [101]. When also oxide traps with a thermally activated barrier are considered in addition to interface states, the OFIT results do not show fast initial degradation or fast recovery of interface states after one second anymore. The real short-term behavior will be elaborately examined in Chapter 6 and 7.

Chapter 6

Short-Term NBTI

During long-term stress, most measurements indicate that ΔV_{TH} follows a power-law as At^n [7, 49, 102, 103]. However, log-like behavior, in particular at short times, has also been reported [11, 40, 84]. Both cases are depicted in Fig. 6.1.

The conventional explanation of the resulting degradation uses elastic hole trapping due to tunneling carrier exchange with the substrate (initial degradation) [11, 84] and the creation of interface states (long-term degradation) [94, 104]. While [94, 104] claim that processes in the short-time scale show a negligible temperature dependence, the latest results support a thermally activated tunneling mechanism [78, 98, 105] (cf. Chapter 5).

Previous short-time measurements using conventional parameter analyzers with a time resolution in the millisecond regime indicate that at least for up to medium stresses a logarithmic time dependence is observed during the first three decades (1 ms up to 1 s) [106]. This logarithmic short-term

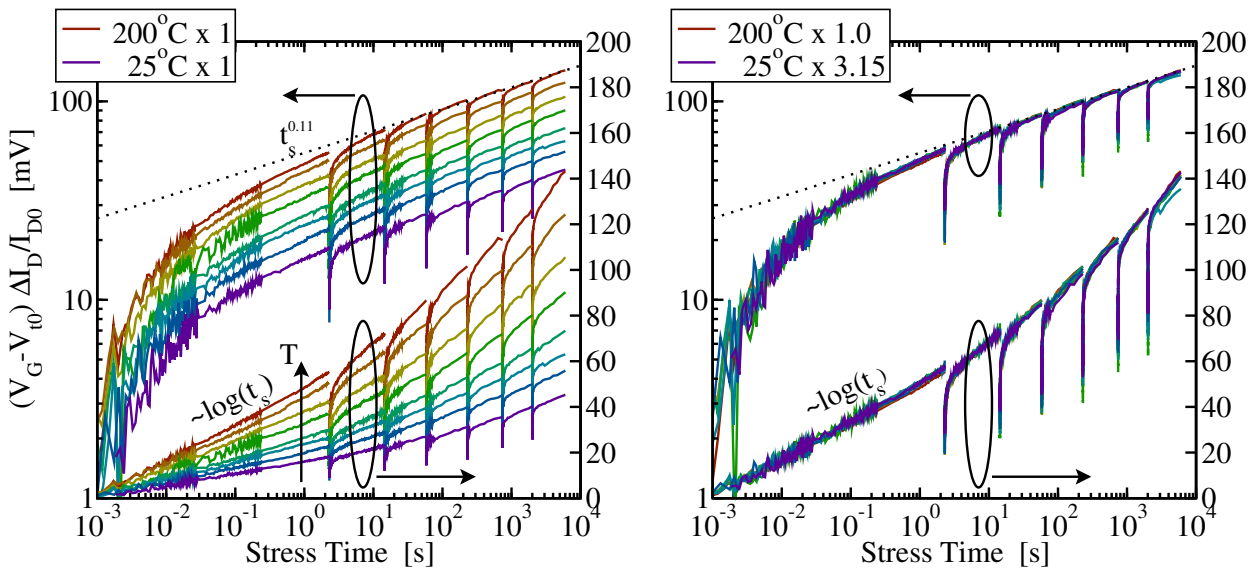


Figure 6.1: A pMOS with $t_{ox} = 1.4$ nm SiON yields a different stress behavior when comparing the short-term to the long-term behavior and the question arises whether there are two mechanisms contributing to NBTI or one. **Left:** Unscaled stress and restress phases of the extended MSM sequence provided by Kaczer *et al.* [17, 18]. **Right:** Scaling the data shown left gives a universal curve which more clearly reveals the $\log(t)$ versus the power-law dependence.

degradation shows a strong temperature activation and a quadratic stress field dependence ($\sim E_{\text{ox}}^2$) up to medium stress ($\approx 5 \text{ MV/cm}$). For longer stress times and higher stress fields ($\approx 7 \text{ MV/cm}$), degradation starts to deviate from the logarithmic behavior [98, 106].

To better understand the underlying mechanisms of short-term NBTI degradation, an extensive study of the short stress time behavior far below the range of milliseconds to seconds needs to be performed. Unfortunately, accurate measurements in these time scales are difficult to access due to noise [42, 107]. In particular, the noise in the μs regime makes it difficult to extract information on the smallest time-constants contributing to the degradation. The currently used measurement methods for fast NBTI evaluation [12] are briefly summarized, based on Chapter 2:

(i) The fast- V_{TH} method [11, 15] (Chapter 2.1.2) interrupts the stress (μs delay) to quickly record V_{TH} during recovery.

(ii) The fast- I_{D} method [17, 18, 20, 30, 106] (Chapter 2.1.1) monitors the drain current I_{D} near V_{TH} , which is then converted to ΔV_{TH} [106] using an initial $I_{\text{D}}(V_{\text{G}})$ curve. This characteristic is only recorded around V_{TH} so as not to prestress the device.

(iii) The on-the-fly (OTF) method [6, 28, 36] records the degradation during stress and hence does not introduce unwanted recovery, but suffers from mobility degradation, which leads to a spurious ΔV_{TH} [41, 108] (Chapter 2.3).

While OTF suffers from the problem of the initial reference measurement, which already stresses the device, the fast- V_{TH} and the fast- I_{D} methods can record an unstressed reference value but suffer from the delay during measurement [12, 18]. Due to its non-stop recording nature, methods (i) and (ii) [12, 18] can continuously monitor recovery and, thus, allow an extrapolation back to shorter measuring delays.

Based on this experience fast rectangular gate pulses are used for short-term NBTI degradation in the range of $1 \mu\text{s}$ to 1 s here. Recalling that previously published results [106] only feature a minimum time of 1 ms means that the number of decades in time for short-term degradation is doubled from three to six. This method is called improved fast pulse method (Chapter 2.2.2) and will be explained thoroughly in the next two sections and is finally compared against the fast- V_{TH} method of [15].

6.1 Gate Pulse Settings

In order to automatically perform the required averaging of the recorded I_{D} , rectangular gate pulses were used for short-term NBTI stresses in the range of $1 \mu\text{s}$ to 1 s , as illustrated in Fig. 6.2. Each gate pulse was followed by a 100 times longer recovery sequence which allowed for full recovery of the built up degradation [14].

Consequently, a pulse train with $t_{\text{lead}} = t_{\text{trail}} = 5 \text{ ns}$, a width $t_{\text{W}} = t_{\text{str}}$, and a period of $t_{\text{P}} = 100 t_{\text{str}}$, consisting of N pulses is used. The product Nt_{P} is only limited by the overall measurement time $t_{\text{M}} = Nt_{\text{P}}$. A compromise between the recovery time in-between pulses ($\approx t_{\text{P}}$) to let the device fully recover and a reasonably high N has to be found in order to gain sufficient measurement accuracy through averaging.

Since the oscilloscope uses a linear time scale, but NBTI stress must be assessed on a logarithmic scale spanning at least 3 to 4 decades, the stress time of 1 s had to be split into three intervals, cf. Fig. 6.2 (right). This allows higher time resolution at the beginning of the stress phase and lower resolution at its end. Since the measurement noise decays with the inverse of the time resolution, with the slower sequences a lower averaging number is necessary to achieve a given amplitude

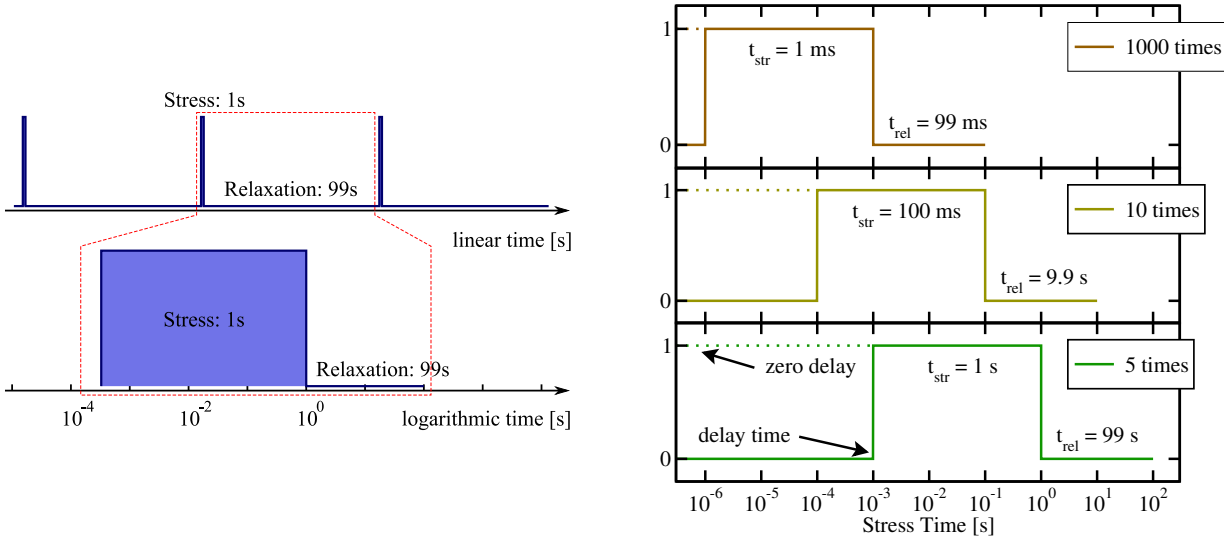


Figure 6.2: Short-term NBTI stress is performed using $1\ \mu\text{s}$ up to 1 s long rectangular gate pulses. **Left:** The very low duty cycle is necessary to achieve full relaxation between stresses. **Right:** The total stress time is split into three sub-intervals including some overlap. To record the same process many times, the above mentioned long recovery time is required. Averaging enhances the amplitude resolution. The number of used pulses is shown in the legend.

Sequence	$t_W = t_{str}$	t_{rel}	t_P	N	Resolution
1	1 ms	99 ms	0.1 s	1000	$0.16\ \mu\text{s}$
2	100 ms	9.9 s	10 s	10	$16\ \mu\text{s}$
3	1000 ms	99 s	100 s	5	$160\ \mu\text{s}$

Table 6.1: Details of the rectangular stress pulses used to maximize the amount of recorded information together with the resolution.

resolution. The according values of t_{str} , t_{rel} , t_P , and N are shown in Tab. 6.1, as well as the resolution, which also equals the minimum stress time of the respective stress sequence.

In order to combine the three sequences into a single degradation curve with a maximum effective resolution from $1\ \mu\text{s}$ to 1 s, the three stress sequences are chosen to overlap for at least one decade of time. Since only differences of currents (I_D) are recorded, the overlap regions provide information to align the sequences to a single stress characteristic. An example is displayed in Fig. 6.3 (left). The offset in $I_D/I_{D,0}$ depends on the different amplification factor in each measurement sequence.

6.2 Data Extraction

Since both the measurement equipment and the pulse generator are operated at their limits, a few points have to be carefully considered during the final data extraction.

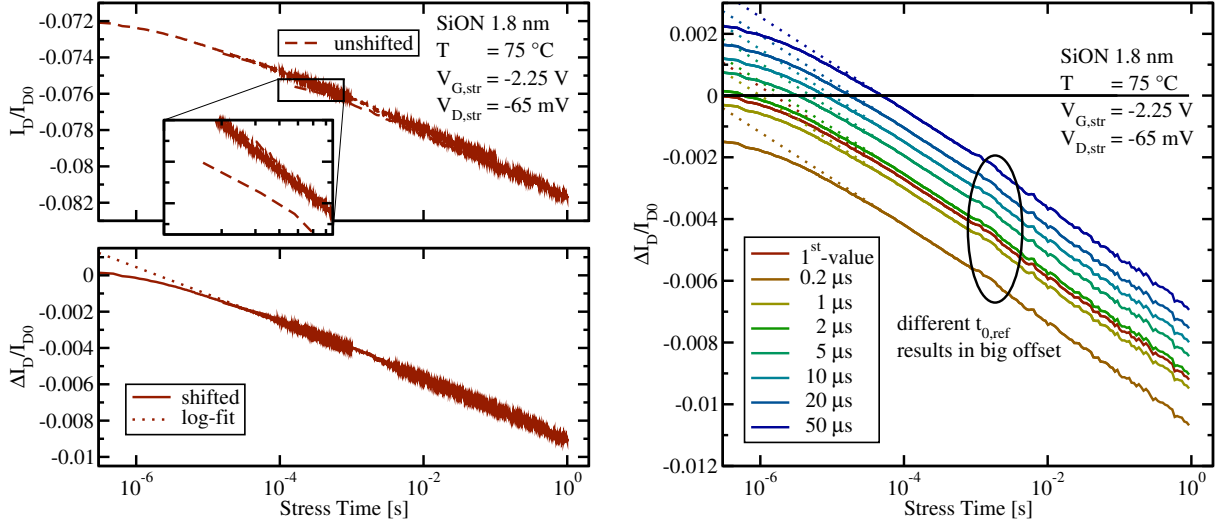


Figure 6.3: Left: Different amplification factors in the DSO settings are responsible for the vertical offset (top). This has to be corrected to make the stress sequences coincide. Merged stress sample (bottom) using a log-fit and shifted to the reference time $t_{0,\text{ref}} = 2 \mu\text{s}$. Right: Different reference times $t_{0,\text{ref}}$ result in different degradation. It can be seen that for $t_{0,\text{ref}} = 50 \mu\text{s}$ about 25% of the $\Delta I_D/I_{D,0}$ are missed. On the other hand, too short $t_{0,\text{ref}}$ are not reasonable and result in a spurious shift by a not-yet steady measurement signal ($t_{0,\text{ref}} = 0.2 \mu\text{s}, 1 \mu\text{s}$). Compare with Fig. 6.4.

6.2.1 Offset

Acquisition of 25 kSamples yields 3 to 4 usable decades in time for each sequence. The combined sequences result in 5 to 6 decades in time, with a possibly too large deviation of $V_{G,\text{str}}$ from the reference $V_{G,\text{str}}^{\text{ref}}$ set at the DSO during the first decade¹. In the remaining decades the data can be either fit by a logarithmic time-dependence

$$\frac{\Delta I_D(t_{\text{str}})}{I_{D,0}} = \frac{I_D(t_{\text{str}}) - I_{D,0}}{I_{D,0}} = -B \log_{10}(t_{\text{str}}/t_{0,\text{ref}}) \quad (6.1)$$

with $I_{D,0} = I_D(t_{0,\text{ref}})$, or a power-law $-A(t_{\text{str}}/t_{0,\text{ref}})^n$ with a very small exponent $n \approx 0.04$. $I_{D,0}$ is obtained at stress-level with a delay $t_{0,\text{ref}}$ and thus not equal to $I_D(0)$ [40] and results in an offset of the relative degradation, see Fig. 6.3 (right).

6.2.2 Initial Measurement as Reference

Unfortunately, the transition from the end of stress to the following recovery is always accompanied by some delay and finite transition times. Effects faster than $1 \mu\text{s}$ are not visible in the experiments. The delay of the first measurement point serving as an initial reference is often discussed in literature [15, 30, 40, 109]. Some authors [42, 102, 104] argue $1 \mu\text{s}$ to be sufficiently short. However, while a reference time $t_{0,\text{ref}}$ taken at 0s would be the ideal case, the real $t_{0,\text{ref}} > 0\text{s}$ always depends on the used equipment. Furthermore, different $t_{0,\text{ref}}$ strongly influence the following stress behavior, cf. Fig. 6.3 (right).

¹The reason why this happens and its consequences will be explained in more detail in Chapter 6.2.3.

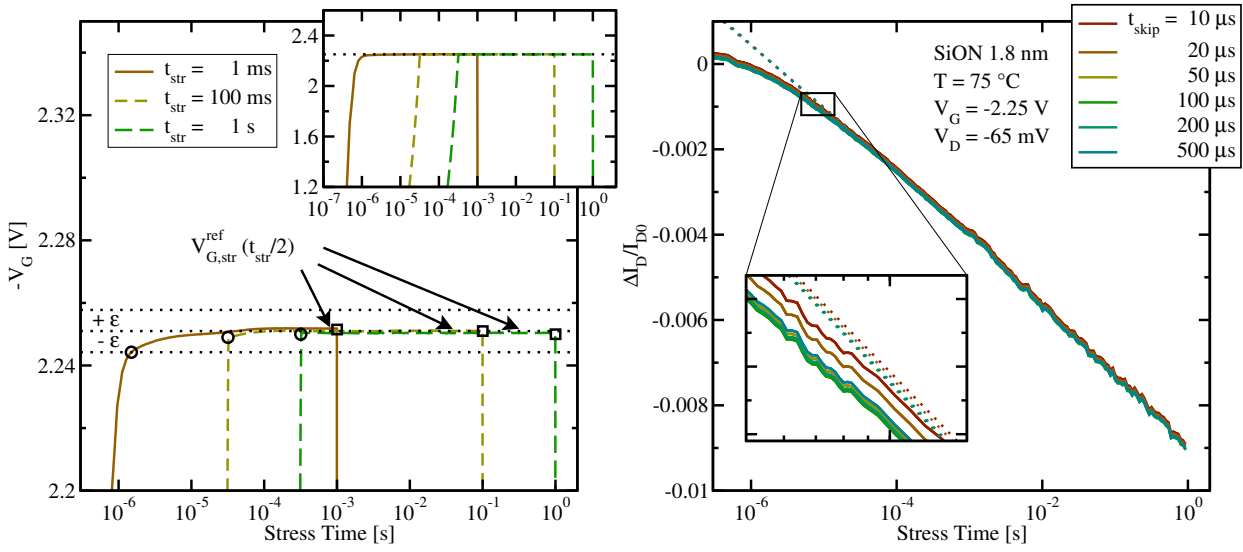


Figure 6.4: **Left:** The main graph is enlarged to make the transient and the overshoot of different stress pulses visible which are shown in the inset. This is due to the limited switching speed of the pulse generator when moving from $V_{G,rel}$ to $V_{G,str}$ and back. The employed error criterion $|(V_{G,str} - V_{G,str}^{ref})/V_{G,str}^{ref}| \leq \epsilon$ is displayed for $\epsilon = 0.3\%$. The first (last) proper values of the pulse for each sequence are marked by circles (squares). The noise is apparent in all three sequences and limits ϵ to extremely small values. **Right:** Logarithmically weighting in time and skipping the first data points corresponding to a parameter t_{skip} does affect the shifting stability but only slightly changes the shift $\Delta I_D/I_{D,0}$ (1%).

6.2.3 Gate Voltage Criteria

In this section it is demonstrated that, in general, fast NBTI measurements have to be taken with a grain of salt. This is largely due to difficulties with synchronization between the stimulus and the actual measurement. So even when the experiment is free of systematic synchronization errors, i.e. switching of the gate voltage and recording of I_D start at the same time, the finite settling time of real signals makes ex-post time zero adjustments necessary. Hence, the time evolution of the actual waveform has to be checked carefully [18]. It turned out that the pulse length is around 0.3% longer than originally set by the pulse generator. This factor has to be accounted for and the real stress times t_{str} of the sequences need to be extracted using the applied gate pulse. As shown in Fig. 6.4 the pulse is affected by the transient behavior and a possible overshoot due to the non-instantaneous switching between $V_{G,rel}$, which is applied in-between the pulses, and $V_{G,str}$. Therefore, after the transition regime, a steady state value of $V_{G,str}$ is determined and set as $V_{G,str}^{ref}$ (usually taken at $t_{str}/2$). Then an error criterion, i.e. $|(V_{G,str} - V_{G,str}^{ref})/V_{G,str}^{ref}| \leq \epsilon$ is employed. Since noise is apparent in all three sequences, ϵ has to be chosen large enough to not disrupt the pulse, usually in the range of $\epsilon \approx 0.3\%$. Starting at $t_{str}/2$ and moving as well to lower (to the beginning of the pulse) and higher (to the end of the pulse) times sets new borders of our accepted stress time t_{str} .

The treatment of the relaxation phase is more complex. It is argued that the noise level is the same during stress and relaxation (the DSO continuously records, using the same settings), and the settling time of the pulse generator in theory is equal regardless if switching from $V_{G,rel}$ to $V_{G,str}$ or vice versa occurred. The criterion for the relaxation phase could then be established as ‘all points extending to both sides of $t = 2t_{str}$ that fulfill $|(V_G - V_{G,rel}^{ref})/V_{G,rel}^{ref}| \leq \epsilon$ ’. This effectively uses the

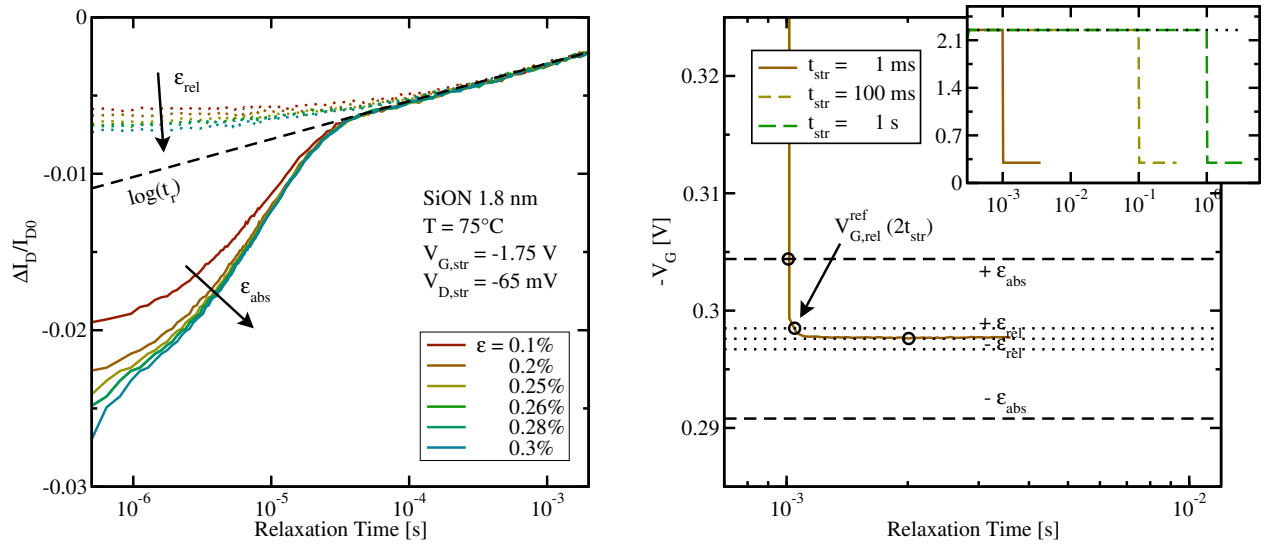


Figure 6.5: **Left:** The extracted change in I_D for different values of ϵ of the relative and absolute truncation criterion is depicted. The logarithmic dependence for longer relaxation times is indicated, too. **Right:** The main graph is enlarged to make the transient and the overshoot visible. The bounds due to both ‘ ϵ_{abs} -criterion’ and the ‘ ϵ_{rel} -criterion’ are displayed for $\epsilon = 0.3\%$, with the first points of the relaxation pulse (after $t_{\text{str}} = 1$ ms) marked by circles.

same *absolute* allowed deviation from $V_{G,\text{rel}}^{\text{ref}}$ as was used during determination of the stress phase, hence this method will be referred to as the ‘ ϵ_{abs} -criterion’. On the other hand, the relative error in I_D (and hence in ΔV_{TH}) that would erroneously be attributed to NBTI is given by the *relative* deviation of V_G , asking for a criterion $|(V_G - V_{G,\text{rel}}^{\text{ref}})/V_{G,\text{rel}}^{\text{ref}}| \leq \epsilon$. This method, which is tighter by a factor of $|V_{G,\text{str}}/V_{G,\text{rel}}| \approx 7$, is referred to as the ‘ ϵ_{rel} -criterion’. Both methods were investigated thoroughly, and the relative method was chosen.

6.2.4 Brute-Force Truncation of the Transient

A second possibility to determine t_{str} is to skip the first data points during the transient until a specific time t_{skip} . This method, displayed in Fig. 6.4 (right), is far easier to implement and gives stable results for various values of t_{skip} . Unfortunately, t_{skip} has to be adjusted manually for every measurement. Hence, the first method is chosen.

6.2.5 Final Setting of Parameters

The finally extracted data is more or less sensitive to the values of the parameters $t_{0,\text{ref}}$ and ϵ . For ϵ a value of 0.3% is used for the stress case, while for relaxation larger values have to be chosen to account for the instantly beginning relaxation. This issue will be dealt with in Chapter 6.5. As can be seen in Fig. 6.4 a $t_{0,\text{ref}}$ slightly after the first value should be selected to both eliminate the influence of the first noisy points and delay time. Hence, $t_{0,\text{ref}} = 2\ \mu\text{s}$ appears a reasonable compromise.

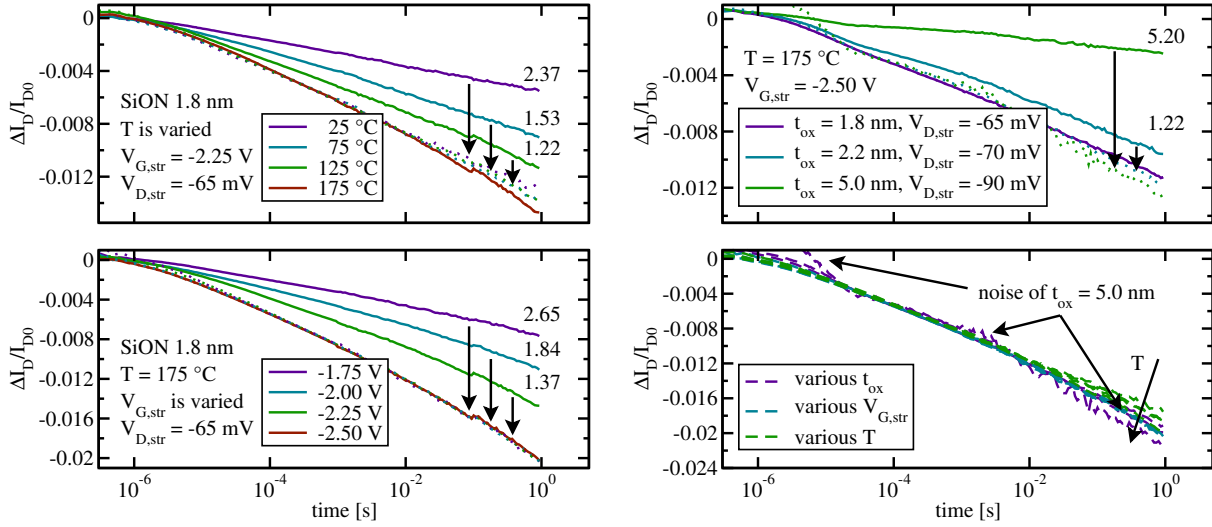


Figure 6.6: Left: The temperature (25 °C, 75 °C, 125 °C, and 175 °C) and voltage dependence (−1.75 V, −2.00 V, −2.25 V, and −2.50 V) of $\Delta I_D/I_{D,0}$ degradation. Scaling to the dotted lines works perfectly for various stress voltages at equal temperatures, while different temperatures lead to a small deviation for $t_{str} > 10$ ms. The scaling factors are also given. Right: $\Delta I_D/I_{D,0}$ for different oxide thicknesses (1.8 nm, 2.2 nm, and 5.0 nm) can be scaled as well. Only the thick device is affected by noise due to the low degradation. The graph at the very bottom combines the three dependencies.

6.3 Logarithmic Stress Behavior

In order to understand the microscopic physics behind the short-time degradation, the temperature, voltage, and oxide-thickness dependence of the prefactor B of (6.1) is investigated. Therefore, a large dataset of stress measurements is collected and analyzed.

6.3.1 Used Samples and Stress Conditions

pMOSFETs from a standard 90 nm CMOS process with plasma-nitrided oxide (around 6% of nitrogen) were used. Two thin oxide devices ($t_{ox} = 1.8$ nm, 2.2 nm) with geometry $W/L = 10 \mu\text{m}/0.12 \mu\text{m}$ and one thicker oxide device ($t_{ox} = 5$ nm) with $W/L = 10 \mu\text{m}/0.24 \mu\text{m}$ were used. The devices were stressed with gate voltages $V_{G,str}$ of −1.75 V, −2.00 V, −2.25 V, and −2.50 V at temperatures of 25 °C, 75 °C, 125 °C, and 175 °C.

6.3.2 Temperature Scaling

The temperature dependence of $\Delta I_D/I_{D,0}$ is displayed in Fig. 6.6 for the thinnest device ($t_{ox} = 1.8$ nm) with $V_{G,str} = -2.25$ V. In the range 25 °C to 125 °C, the data can be perfectly fit by a logarithmic time dependence (differences would not even be visible in the plots). A slight deviation is observed for higher temperatures for $t_{str} > 10$ ms, possibly due to the onset of the mechanism responsible for the long-time power-law behavior with a larger power-law exponent $n \approx 0.12$. This might be due to the dependence of two compound power-laws which are discussed in literature [49]. Apart from that, different temperatures can be scaled well to the data at $T_{ref} = 175$ °C, as shown by the dotted lines in Fig. 6.6, and the indicated scaling factors marked by arrows.

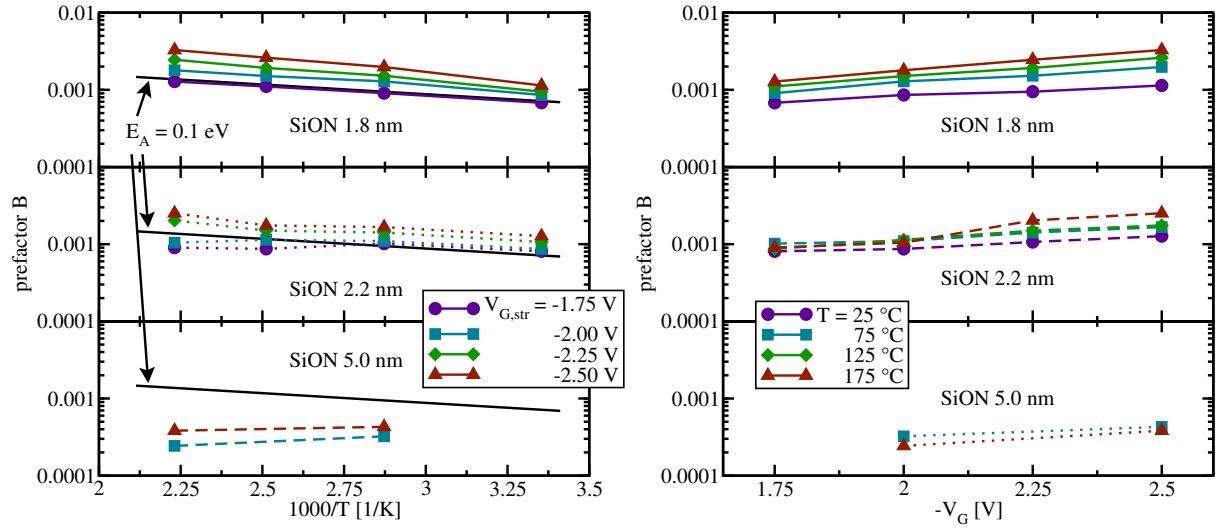


Figure 6.7: **Left:** Arrhenius plot of the prefactor B of the log-fit, extracted from three different t_{ox} for different $V_{G, \text{str}}$. An activation energy E_A of about 0.1 eV is obtained for $t_{\text{ox}} = 1.8$ nm and $t_{\text{ox}} = 2.2$ nm, represented by the black solid line. Degradation for the $t_{\text{ox}} = 5.0$ nm devices was too noisy due to too low $E_{\text{ox}} \sim (V_{G, \text{str}} - V_{\text{TH}})/t_{\text{ox}}$. Scale is equal for all plots. **Right:** Prefactor B of the log-fit plotted for different t_{ox} with different temperature T . While $t_{\text{ox}} = 1.8$ nm shows a clear temperature activation, $t_{\text{ox}} = 5.0$ nm does not due to the low electric stress field. For $t_{\text{ox}} = 2.2$ nm the transition of the temperature dependence is visible at $T = 175$ °C between $V_{G, \text{str}} = -2.00$ V and $V_{G, \text{str}} = -2.25$ V.

6.3.3 Voltage Scaling

The voltage dependence is depicted for $t_{\text{ox}} = 1.8$ nm and $T = 175$ °C (Fig. 6.6). Scaling to $V_{G, \text{str}}^{\text{ref}} = -2.50$ V leads to perfect congruence. Again, the scaling factors are shown next to their corresponding traces.

6.3.4 Oxide Thickness Scaling

Due to the relatively low $\Delta I_D/I_{D,0}$ degradation for $t_{\text{ox}} = 5.0$ nm resulting from the low-voltage stress conditions studied here (small E_{ox}), noise seriously limits the accuracy. Nonetheless, good scalability for different t_{ox} devices (1.8 nm, 2.2 nm, and 5.0 nm) can be obtained (Fig. 6.6).

6.3.5 Extracted Prefactors

The prefactor B of the log-fit for various t_{ox} , $V_{G, \text{str}}$, and T is displayed in Fig. 6.7. In agreement with previous experiments, it is observed that low $V_{G, \text{str}}$ results in small temperature activation, while $V_{G, \text{str}}$ larger than the operating voltage of the MOSFET gives a notable activation energy of 0.1 eV. Note that this value is in agreement with activation energies extracted at long stress times [106]. Fitting the data to a power-law $A(t_{\text{str}}/t_{0, \text{ref}})^n$ results in an exponent $n \approx 0.04$ for short-term stress, roughly a third of the often reported $n \approx 0.12$ of the long-term behavior. This is in very good accordance with the standard E_A for NBTI stress and accounts for a strong $V_{G, \text{str}}$ dependence, excluding elastic hole tunneling.

The right graph of Fig. 6.7 represents the prefactor B plotted for different t_{ox} at different temperatures. In the devices with $t_{\text{ox}} = 1.8$ nm, all the stress voltages are above the operating voltage and result in a marked temperature activation. For $t_{\text{ox}} = 2.2$ nm the transition from no temperature activation to temperature activation is observed between $V_{\text{G, str}} = -2.00$ V and $V_{\text{G, str}} = -2.25$ V for $T = 175$ °C. For the thickest oxides used in this study, $t_{\text{ox}} = 5.0$ nm, the applied stress fields are too small to lead to a meaningful degradation². Therefore no objective statement can be made on temperature activation concerning the here presented devices with $t_{\text{ox}} = 5.0$ nm .

However, the experiments performed on devices with smaller oxide thicknesses support thermally activated tunneling mechanism [98] rather than elastic (and thus temperature-independent) hole tunneling [94].

6.4 Power-Law Stress Behavior

In contrast to Chapter 6.3, where the degradation of the drain current is directly fit by (6.1), the drain current is now first converted to an approximate threshold voltage shift using the simple OTF1 relation derived in Appendix A.1

$$\Delta V_{\text{TH}}(t_{\text{str}}/t_{0,\text{ref}}) \approx \frac{I_{\text{D}}(t_{\text{str}}) - I_{\text{D}}(t_{0,\text{ref}})}{I_{\text{D}}(t_{0,\text{ref}})} (V_{\text{G}} - V_{\text{TH},0}) \quad (6.2)$$

$$= \frac{\Delta I_{\text{D}}(t_{\text{str}})}{I_{\text{D}}(t_{0,\text{ref}})} (V_{\text{G}} - V_{\text{TH},0}). \quad (6.3)$$

Note that $I_{\text{D},0}$ is obtained at stress-level with a delay $t_{0,\text{ref}}$ and is thus *not* equal to $I_{\text{D}}(0)$ [40], resulting in an offset of the relative degradation. Also, the conversion (6.3) ignores any potential degradation in the mobility and is thus affected by an as-of-yet unknown error [41, 108]. The threshold voltage is extracted at $I_{\text{D}} = 70$ nA · W/L, which yields $V_{\text{TH}} \approx -0.3$ V. Then ΔV_{TH} is fit by

$$\Delta V_{\text{TH}}(t_{\text{str}}/t_{0,\text{ref}}) \approx B \log_{10}(t_{\text{str}}/t_{0,\text{ref}}) + C. \quad (6.4)$$

In order to circumvent issues with the logarithmic fit caused by offset data due to the uncertainty in $I_{\text{D},0}$, the parameter C is included. Besides, it is tried to fit the data to a power-law of the form

$$\Delta V_{\text{TH}}(t_{\text{str}}/t_{0,\text{ref}}) \approx A (t_{\text{str}}/t_{0,\text{ref}})^n + D. \quad (6.5)$$

Again, the parameter D is introduced to account for the offset in $I_{\text{D},0}$.

Interestingly, it turns out that the logarithmic fit (6.4) is always possible, while the power-law fit (6.5) produces reasonable results for high temperatures and high $V_{\text{G, str}}$ only. In that high-stress regime, power-law exponents around 0.04 are obtained. For weaker stresses, the exponent n in (6.5) tends towards zero, which corresponds to a first-order Taylor expansion of (6.5) on a logarithmic scale. As such, in this regime the power-law fit (6.5) becomes equivalent to the logarithmic fit (6.4).

This behavior is illustrated in Fig. 6.8 (top). The data obtained from the harshest stress conditions ($V_{\text{G, str}} = 2.50$ V, $T = 175$ °C, and $t_{\text{ox}} = 1.8$ nm) gives a stable fit with $n = 0.041$. For the other extreme case ($V_{\text{G, str}} = 1.75$ V, $T = 25$ °C, and $t_{\text{ox}} = 1.8$ nm) the fitting algorithm gives an exponent n of practically zero. For the case of the non-converging exponent n the logarithmic and power-law fits coincide.

²To account for that the degradation should better be referred to the oxide electric field $(V_{\text{G, str}} - V_{\text{TH}})/t_{\text{ox}}$ instead of the stress voltage.

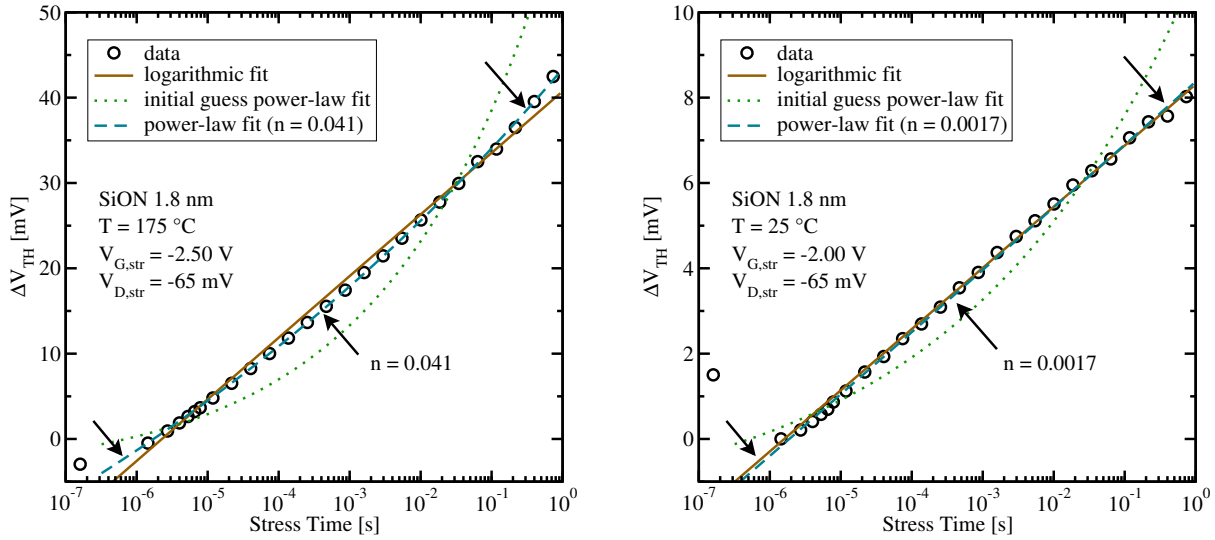
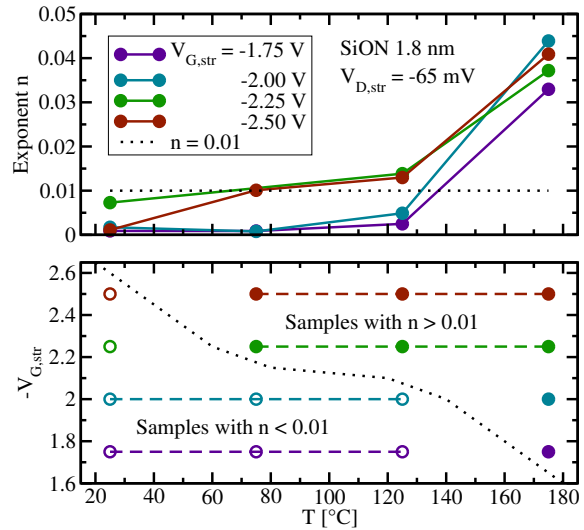


Figure 6.8: **Top Left:** At the highest stress condition ($V_{G,\text{str}} = -2.50 \text{ V}$, $T = 175 \text{ }^\circ\text{C}$) the recorded data slightly deviates from a logarithmic dependence and can be nicely fit using a power-law. **Top Right:** By contrast, data recorded using a lower stress condition ($V_{G,\text{str}} = -2.00 \text{ V}$, $T = 25 \text{ }^\circ\text{C}$) nearly perfectly follows a logarithmic behavior and cannot be properly fitted using a power-law. **Center Right:** Only data recorded during heavier stress yield a reasonable power-law exponent n . **Bottom Right:** Using the (arbitrary) value of $n = 0.01$ as a threshold criterion, a high-stress region, where a deviation from the logarithmic behavior is observed, can be clearly identified.



Consequently, the power-law fit only makes sense for high temperatures and/or high $V_{G,\text{str}}$, as displayed in Fig. 6.8 (center and bottom right). There, the extracted $n \approx 0.04$ for short-term stress is roughly one third of the often reported $n \approx 0.12$ of the long-term behavior.

6.5 Relaxation Behavior

For relaxation, based on the assumption of full recovery after each stress/relaxation cycle, $I_{D0,\text{rel}} = I_D(t_P)$ is chosen with t_P being the pulse period, and implying $\Delta I_D(t_P) = 0$. Therefore $I_{D0,\text{rel}}$ is independent of ϵ , whereas $I_{D0,\text{str}} = I_D(t_{0,\text{ref}})$ depends on the ϵ used. Note that due to record length constraints of the DSO, not the entire relaxation characteristic up to t_P is recorded, but only the initial relaxation up to around three to four times t_{str} . The point $t = t_P$ nevertheless is available in the pre-trigger data of the DSO.

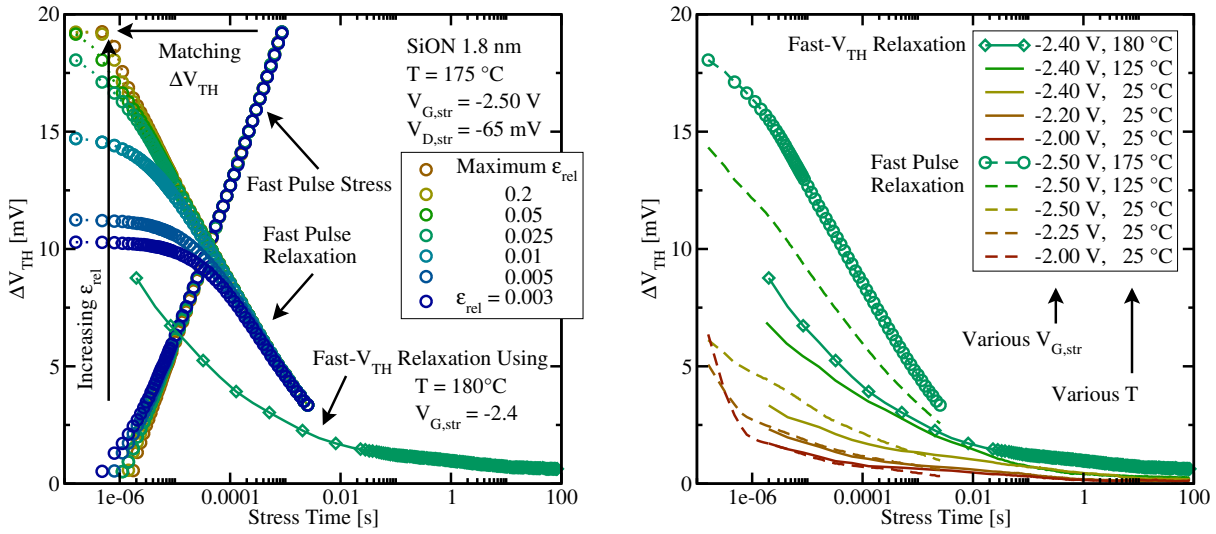


Figure 6.9: Comparison of the fast- V_{TH} measurement and the fast pulsed $I_D(V_G)$ -method both developed by Reisinger [12]. **Left:** Depending on the relative epsilon criterion ϵ_{rel} , more or less points are considered for the relaxation curves leading to different initial slopes. While $\epsilon_{rel} = 0.003$ results in an artificial plateau, as the data points are stretched out towards shorter times, $\epsilon_{rel} = 0.025$ shows good agreement with the data obtained from the fast- V_{TH} measurement, which do not have a plateau. **Right:** Due to a limited number of available devices it was not possible to perform the same stressing conditions using the two different measurement techniques. However, comparable conditions are plotted in the same color to show the quite good agreement of both techniques.

Data extraction turns out to be extremely sensitive to the choice of ϵ , indicating that the settling time of V_G plays a crucial role in OTF experiments. To demonstrate this fact, Fig. 6.9 (left) shows relaxation after $t_{str} = 1$ ms for different values of ϵ . If the criterion is too conservative, i.e. ϵ is chosen small, thereby cutting off the initial relaxation phase, the shape of the relaxation characteristics is significantly altered. On the other hand, too large values of ϵ , i.e. too liberal limits for gate voltage settling, may produce spurious relaxation transients. In other words, with different ϵ_{rel} values more or less points are considered for the relaxation curves leading to different initial slopes. The ‘real’ initial data, as seen in Fig. 6.9, then becomes artificially dispersed when too many data points are cut off (small ϵ).

Assuming the relaxation follows $\log(t_{rel}/t_0)$ as indicated by the red curve in Fig. 6.9, and setting the starting point of the extracted relaxation to later times (through smaller ϵ) gives a dependence of $\log((t_{rel} + \Delta t)/t_0)$, which produces the artificial plateaus seen with the blue curves in the figure. This may lead to the wrong conclusion that the time constants are smaller than they actually are. Possibly the saturation towards smaller relaxation times found in [42, 102, 107] could be explained that way, i.e. the plateaus observed are not a feature of NBTI relaxation but an artifact due to finite settling times and synchronization inaccuracies, in turn invalidating the assertion that a measurement delay in the micro-second regime is sufficient to correctly capture the relaxation characteristics of NBTI. Besides, though an ϵ_{rel} of 0.025 seems to be quite large, the resulting V_G only lies within 7.5 mV of the settled $V_{G,rel}$. Therefore these values well account for the relaxation region.

6.6 Fast Ramp versus Fast- V_{TH} -Method

The previously discussed epsilon criterion in addition to the demonstrated full recovery after each stress pulse enables back extrapolation to the very beginning of the relaxation phase of the fast pulsed $I_{\text{D}}(V_{\text{G}})$ -method. For $t_{\text{str}} = 1$ ms this method and the fast- V_{TH} method are compared including different voltages and temperatures. It has to be mentioned that the temperature dependence of $V_{\text{TH},0}$ was not exactly measured, and hence is missing for the precise ΔV_{TH} extraction and that in addition mobility changes with temperature have been neglected. However, the results shown in Fig. 6.9 (right) still show good qualitative agreement of both measurement techniques.

6.7 Conclusions

Ultra-fast short-time NBTI stress and relaxation measurements from the μs to the seconds regime using different temperatures, stress voltages, and oxide thicknesses have been performed. A large dataset is examined here using well defined extraction parameters ($t_{0,\text{ref}}$, t_{skip} , and ϵ). Amongst them the reference time $t_{0,\text{ref}}$ is identified as the most crucial one. It can be seen that depending on the range used for the data extraction (ϵ) the reference time $t_{0,\text{ref}}$ is also changed. While a settled gate pulse, i.e. a small ϵ , does not contain the full degradation and relaxation data and may therefore indicate a wrong distribution of time constants, too broad limits of ϵ_{rel} may produce spurious relaxation transients due to a limited resolution of smaller than $1 \mu\text{s}$. Comparing the different gate voltage criteria taken for the OTF routine yields that choosing a rather large ϵ_{rel} reflects the completely different fast- V_{TH} measurement method best.

In the initial degradation phase, which is often explained by elastic hole trapping, the data can be well fit by a logarithmic time dependence [12,15,42]. As this log-dependence is considerably distorted during long-term measurements, alternatively a power-law using an exponent considerably smaller ($n \approx 0.04$) than generally observed during long-time stress ($n \approx 0.12$) can be used. However, the main disadvantage of the power-law is that the fit is ill defined for up to medium stress conditions. Only high temperatures and/or high $V_{\text{G,STR}}$ show the aforementioned small n .

Moreover, the extracted activation energy of about 0.1 eV is compatible with the values typically obtained during long-time stress [106]. The temperature and voltage dependencies of stress and relaxation rule out elastic and thus temperature-independent hole tunneling as being responsible for short-time NBTI degradation as proposed by [94,104]. A possible explanation could involve an inelastic tunneling process [98].

Chapter 7

Relaxation of Negative/Positive BTI

As the time constant distribution of the microscopic defects behind BTI turn out to be a key issue, the apparent differences in relaxation behavior of negative and positive BTI (NBTI and PBTI) on pMOSFETs, as depicted in Fig. 7.1, are now examined under that perspective.

Although PBTI on pMOSFETs is not regarded as technologically important as NBTI, it provides a valuable probe of the underlying physical degradation mechanism. The most intriguing observation is that both negative and positive bias stress create positive charges in the oxide [30], which was already demonstrated in Chapter 4.2. However, so far the NBTI and PBTI stress conditions were only compared in a qualitative way, i.e. strong inversion was usually opposed to strong accumulation with undetermined specifications concerning the exact gate voltages or oxide electric fields applied.

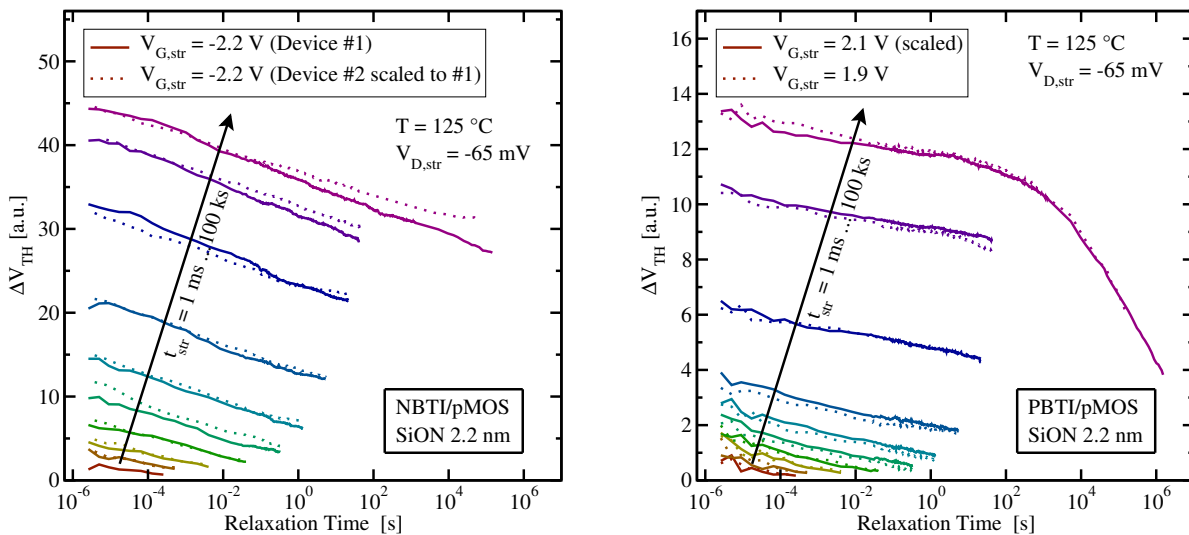


Figure 7.1: While after short stress the relaxation does not show significant differences except slightly varying slopes, the distinct relaxation behavior after NBTI and PBTI is obvious when monitoring the long-term relaxation tail after the last stress sequence. The stress time is increased in steps of one decade with the only exception at 50 ks. Note that degradation data obtained with equal absolute values of the oxide electric field are compared here.

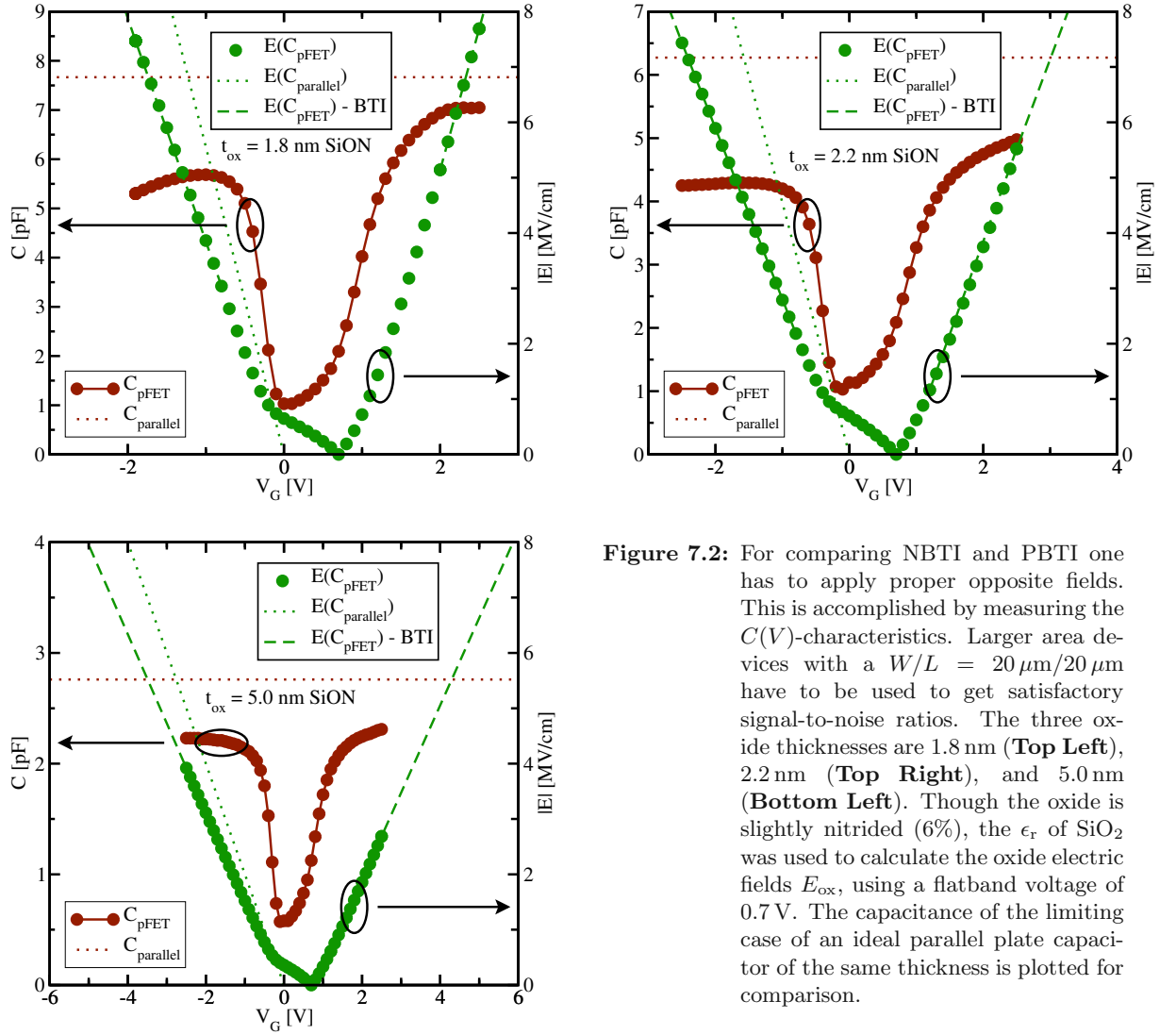


Figure 7.2: For comparing NBTI and PBTI one has to apply proper opposite fields. This is accomplished by measuring the $C(V)$ -characteristics. Larger area devices with a $W/L = 20 \mu\text{m}/20 \mu\text{m}$ have to be used to get satisfactory signal-to-noise ratios. The three oxide thicknesses are 1.8 nm (**Top Left**), 2.2 nm (**Top Right**), and 5.0 nm (**Bottom Left**). Though the oxide is slightly nitrided (6%), the ϵ_r of SiO_2 was used to calculate the oxide electric fields E_{ox} , using a flatband voltage of 0.7 V. The capacitance of the limiting case of an ideal parallel plate capacitor of the same thickness is plotted for comparison.

For a quantitative analysis of the recovery following NBTI and PBTI stress, long stress times t_{str} between 100s and 100ks are essential. The same technology (6%-SiON-pMOSFET) as used in Chapter 6 was compared by the fast- V_{TH} method of [15] using three different oxide thicknesses ($t_{\text{ox}} = 1.8 \text{ nm}, 2.2 \text{ nm},$ and 5.0 nm) and the corresponding geometries of $W/L = 20 \mu\text{m}/0.12 \mu\text{m}, 20 \mu\text{m}/0.12 \mu\text{m},$ and $20 \mu\text{m}/0.24 \mu\text{m}$ at a constant temperature of 125°C . Depending on the oxide thickness the same applied stress voltage causes a totally different oxide electric field. This is due to capacity of the MOSFET with its principle already explained in Chapter 2.6. The resulting electric field at the surface of the semiconductor E_s can be experimentally estimated by using the following relation:

$$E_s(V) = \frac{1}{\epsilon_r \epsilon_0 W L} \int_{V_{\text{fb}}}^V C(V') dV' \quad (7.1)$$

where $C(V)$ denotes the capacity of the MOSFET, V_{fb} the flatband voltage, and W and L the width and length of the device. The $C(V)$ -characteristics and the corresponding electric field are

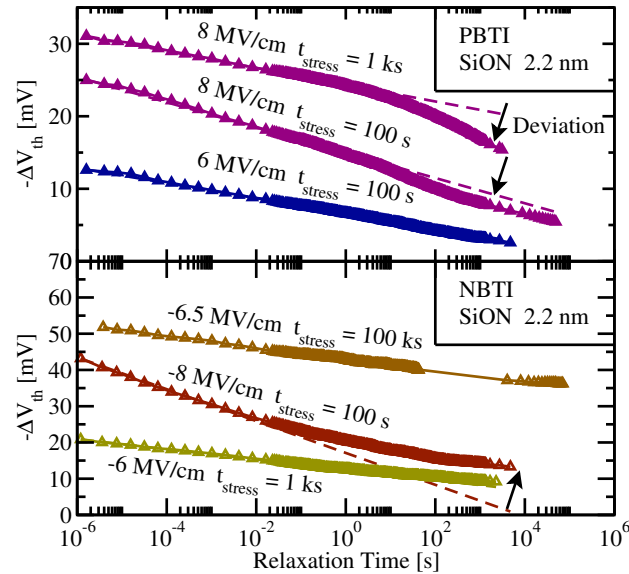


Figure 7.3: Samples with an oxide thickness of 2.2 nm stressed using various NBTI/PBTI-conditions from 100 s up to 10 ks. Depending on the type of stress, there is either no deviation from a logarithmic recovery behavior, a deviation downwards (PBTI) or upwards (NBTI). While for weak NBTI/PBTI-conditions ($E_{\text{ox}} = \pm 6 \text{ MV/cm}$ and $t_{\text{str}} = 100 \text{ s}$) a logarithmic fit of the relaxation is possible, this is not the case for the other heavier stress conditions.

shown in Fig. 7.2 for the different device geometries with a constant flatband voltage of 0.7 V. From this figure it can further be seen that in addition to the nonzero flatband voltage the electric field during NBTI and PBTI is not symmetric. To create comparable degradation conditions (not comparable degradation shifts) for both NBTI and PBTI, the same effective field is of interest, i.e. the same magnitude, but opposite sign. Based on the experimental $C(V)$ -characteristics in Fig. 7.2 the required stress voltage $V_{G,\text{str}}$ can be obtained for both NBTI and PBTI. As an example, to achieve an $E_{\text{ox}} = \pm 6 \text{ MV/cm}$ for $t_{\text{ox}} = 2.2 \text{ nm}$ gate voltages of +2.65 V for PBTI and -2.05 V for NBTI have to be applied.

7.1 Raw Measurement Results

Though only the very last relaxation curves of the MSM-sequence are depicted for various oxide electric fields E_{ox} in Fig. 7.3, with the corresponding $V_{G,\text{str}}$ values obtained from Fig. 7.2, the following similarities of NBTI and PBTI can be summarized: (i) The V_{TH} -shift is always negative, apparently due to positive charge build-up during stress. (ii) Up to medium stresses the degradation also recovers in a similar fashion. (iii) Both NBTI and PBTI show nearly perfect logarithmic relaxation when stressed up to $\pm 6.5 \text{ MV/cm}$ for 100 ks, yielding a constant recovery rate per decade.

However, there are two main differences between NBTI and PBTI stress and recovery: (i) The degradation during PBTI stress is about a factor two smaller than that built up during NBTI. (ii) Deviations are found when comparing the two cases of low field ($E_{\text{ox}} = \pm 6 \text{ MV/cm}$) versus high field ($E_{\text{ox}} = \pm 8 \text{ MV/cm}$), which is emphasized in Fig. 7.3. For NBTI it appears that the strong relaxation in the initial phase ranging from $1 \mu\text{s}$ to about 100 ms slows down to finally saturate.

The saturation level was already defined as permanent component (in contrast to the recoverable component) [29, 30]. After [110] this permanent component follows a power-law. In contrast, for high-field PBTI stress the recovery is first delayed and then pronounced. The relaxation curve here has an S-shape, which is observed for the first time, because it is obviously only visible for long relaxation times.

7.2 Schematic Recovery Behavior

The question arises whether this different recovery shapes are due to an artifact again (cf. Chapter 6.5) or not. Therefore, the key findings are presented first. Given that aid it is possible to discuss the experimental results afterwards.

Let us assume a recovery trace that shows both features, early as well as delayed recovery. Considering the assumptions that no recovery is missed in the beginning and no additional negative charges are created at the same time till the total charges are released again yields the complete recovery trace after BTI stress, schematically depicted in the top left of Fig. 7.4. Unfortunately the full features of the recovery after typical BTI stress are rarely visible, cf. the curve of $E_{\text{ox}} = -8 \text{ MV/cm}$ and $t_{\text{str}} = 100 \text{ s}$ in Fig. 7.3, as often only a part of the S-shaped recovery characteristic can be recorded by the experiment.

While for PBTI only the upper section of the whole relaxation curve is visible, it is the lower section for NBTI. Within these sections the curvature marks the transition between the initial and the late phase of the recovery respectively. By using this curvature to detect a change of the relaxation the recovery following PBTI versus NBTI stress is now analyzed in more detail.

7.3 Extraction Routine

The determination of the curvature following bias temperature stress is displayed in the left of Fig. 7.5. First, each relaxation of V_{TH} is referred to its initial $V_{\text{TH},0}$ and is plotted as ΔV_{TH} as a function of $\log(t)$. The first decades as well as the last decade in time are used to fit the experimental data with a logarithm of the form $a + b \log(t)$, giving the initial and long term recovery behavior. Eventually, the intersection of the two fits results in the “kink points” τ_{A} and τ_{B} . While τ_{A} is used to describe the initial recovery phase generally observed after PBTI, τ_{B} is used for the long term recovery as observed after NBTI. These two cases are depicted in the right of Fig. 7.4. However, the kink-point-method does not work properly with too similar logarithmic prefactors b_1 and b_2 due to glancing intersection, compare $E_{\text{ox}} = \pm 6 \text{ MV/cm}$ and $t_{\text{str}} = 100 \text{ s}$ in Fig. 7.5 (left). For the already discussed complete recovery trace with its S-shape, the first and second fit become nearly parallel resulting in an undetermined kink point.

7.4 Discussion of the Experimental Output

Before the previously described extraction routine is applied on the experimental output of various oxide thicknesses, the origin of the curvature is described schematically for the 1.8 nm thick oxides first. On the basis of Fig. 7.5 (left) each relaxation curve apparently contains two contributions, one depending on the stress time t_{str} and one depending on the oxide electric field E_{ox} .

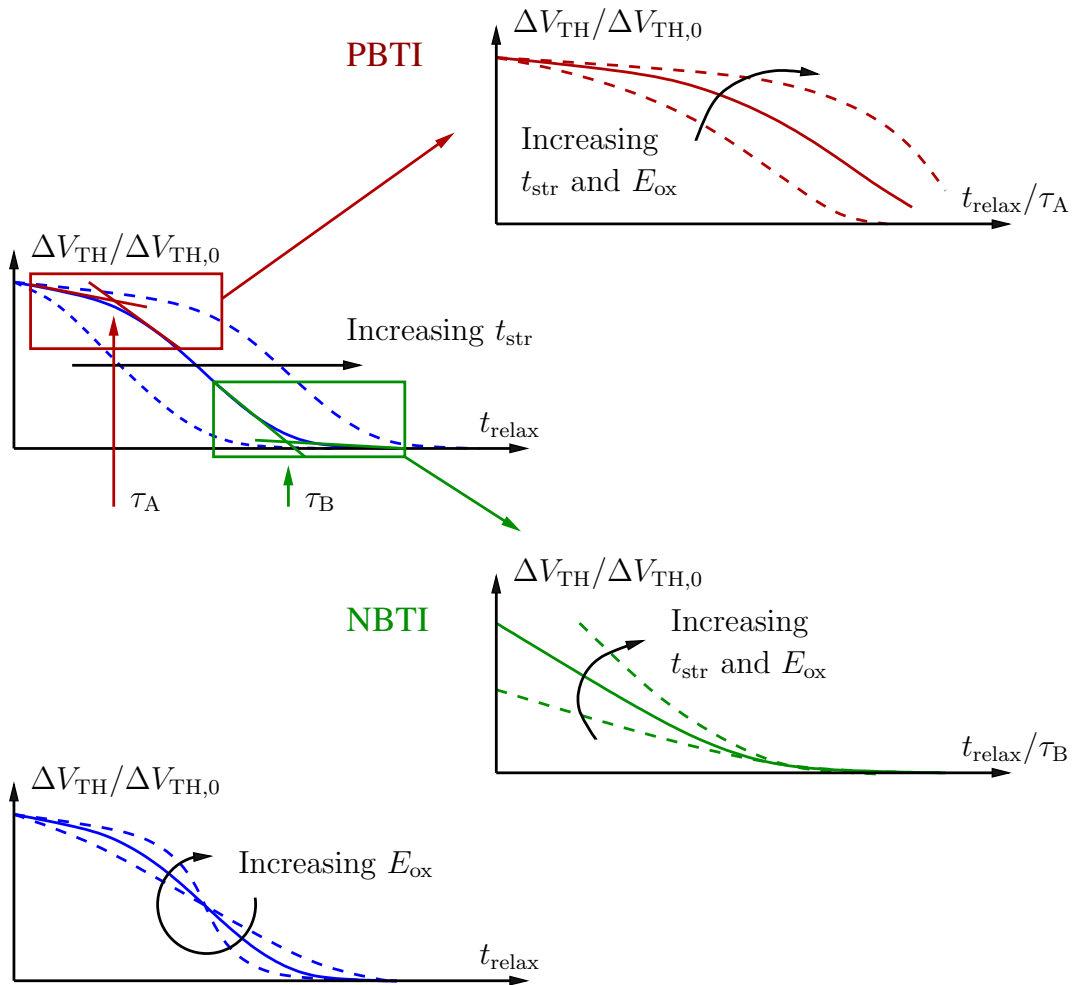


Figure 7.4: A schematic recovery trace after bias temperature stress is shown as solid line. **Top Left:** The full S-shape is only observable under certain conditions, e.g. $E_{ox} = -8 \text{ MV/cm}$ and $t_{str} = 100 \text{ s}$. For longer t_{str} the whole curve is shifted to higher t_{rel} . PBTI mainly shows the characteristics in the top left (red) box, whereas the behavior after NBTI stress typically proceeds as shown in the bottom right (green) box. Within these sections τ_A and τ_B depend on the curvatures and mark the transition between the initial and the concluding phase of the recovery. **Bottom Left:** In combination with the oxide electric field dependence, the behavior of PBTI and NBTI can be obtained by scaling with τ_A and τ_B . **Top Right:** Increasing stress conditions (t_{str} and/or E_{ox}) for PBTI yield smaller relaxation rates per decade at earlier t_{rel}/τ_A , followed by larger relaxation rates afterwards. When extending the observation period towards larger t_{rel}/τ_A , the transition back to smaller relaxation rates becomes visible. **Bottom Right:** Increased stress conditions after NBTI feature increased relaxation up to τ_B . Extending the observation period towards smaller t_{rel}/τ_B after NBTI stress is often not possible due to the limited measurement speed.

7.4.1 Stress Time Component

First, the position of the curvature is characterized by τ_A for PBTI and τ_B for NBTI after Fig. 7.4. τ_A and τ_B obviously depend on the stress conditions, which is displayed in Fig. 7.5 (right) in more detail. The longer the device is stressed for equal E_{ox} , the later the kink occurs. Interestingly the kink-time normalized to the stress time t_{str} is roughly a constant, indicating a connection between

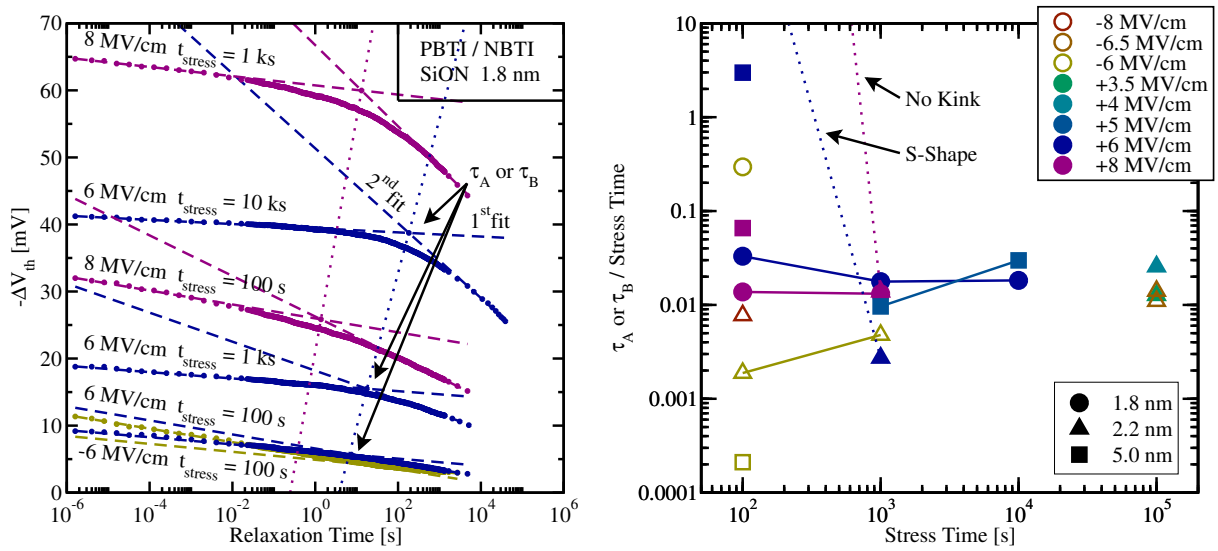


Figure 7.5: Left: The relaxation behavior starts to deviate from the logarithmic shape at harsher stress conditions. By fitting the beginning and the end of the recovery traces separately, a kink point τ_A or τ_B at the extrapolated intersection of the fits can be obtained, which characterizes the curvature. Comparing kink points of the same electric field and oxide thickness for different stress times (connected by dotted lines) shows that this curvature is stronger at longer stress times and delayed with time. Right: As already shown by the dotted lines in the left of Fig. 7.5, a longer stress increases the kink time τ_A for PBTI or τ_B for NBTI. Scaling τ_A or τ_B to t_{str} reveals a proportionality of the kink time and t_{str} . The extraction does not work properly with two kinks at τ_A and τ_B or without a kink, due to the glancing intersection of the fits, resulting in an ambiguous kink point. The positions for these special cases at $t_{str} = 100$ s are indicated by dotted lines.

t_{str} and the recovery behavior (cf. the universal recovery in Chapter 4.1). This is also schematically shown in Fig. 7.4, where increasing t_{str} shifts the relaxation curve towards larger t_{rel} .

7.4.2 Oxide Electric Field Component

To be able to discuss the effect of E_{ox} on the relaxation we again refer to Fig. 7.5 (left). When comparing the devices stressed for 1 ks with 6 MV/cm and 8 MV/cm, τ_A has approximately the same value, but the relative relaxation with respect to its very different $\Delta V_{TH,0}$ changes as depicted in Fig. 7.4 (bottom left). With a raise in the oxide electric field the recovery sets in late, but proceeds faster. This effect is now explored by specifying the slopes of the first (b_1) and second (b_2) logarithm, i.e. the short-term and long-term relaxation behavior.

7.5 Short-Term and Long-Term Relaxation

As illustrated in Fig. 7.6 (left) the initial relaxation rate b_1 after NBTI stress is higher than its PBTI counterpart. For NBTI b_1 increases with increasing E_{ox} , while for PBTI b_1 only slightly increases with increasing E_{ox} . Due to also higher $\Delta V_{TH,0}$ at the beginning of the relaxation with higher E_{ox} , the effect even results in lower relative recovery per decade with higher E_{ox} . Furthermore, for PBTI b_1 decreases with increasing t_{str} because of the higher contributing permanent part.

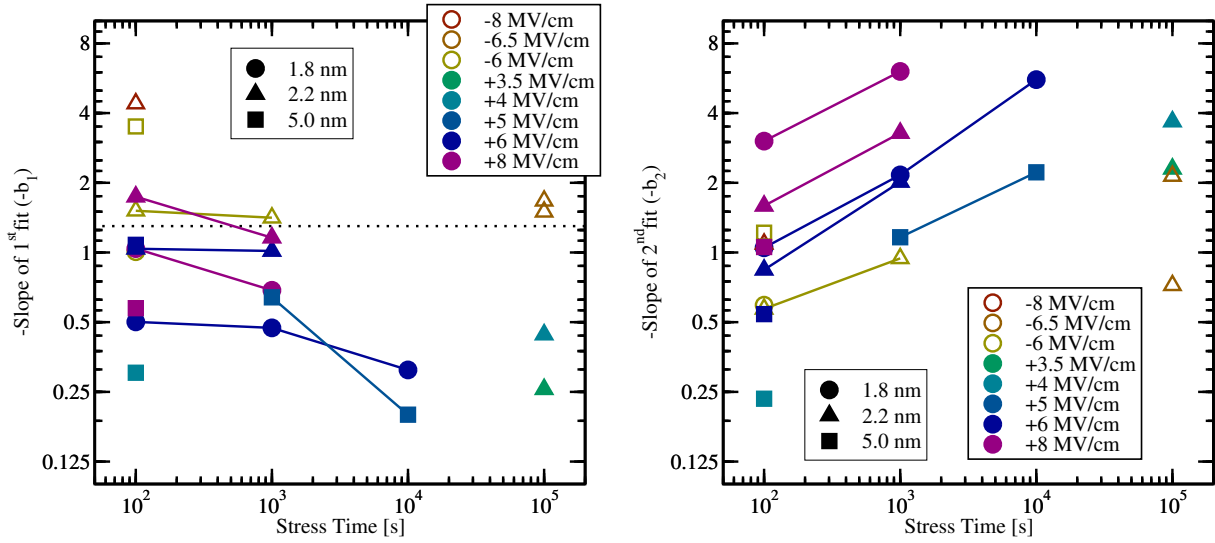


Figure 7.6: The slopes of the first (b_1) and second (b_2) logarithm are plotted over t_{str} as a function of the device thickness t_{ox} and its oxide electric field E_{ox} . **Left:** The slope b_1 of NBTI stress is higher than that resulting from PBTI stress, and decreases with t_{str} as shown by the solid lines. The dashed line denotes the boundary between NBTI and PBTI. The initial relaxation slopes b_1 increase from 0.196 (PBTI) to 4.35 (NBTI), which is a factor of more than 20. This demonstrates the different initial relaxation behavior following NBTI and PBTI stress. **Right:** The slope b_2 for the long-term characteristics increases with t_{str} and E_{ox} and clearly reflects the increased relaxation after PBTI stress with values ranging from 0.53 (NBTI) to 5.99 (PBTI). Combining this fact and recalling that PBTI does practically not recover during the first few seconds supports the assumption that the performed kind of stress condition already constitutes the short-term and long-term relaxation.

In contrast, the long-term relaxation b_2 increases with t_{str} and E_{ox} , which clearly shows enhanced relaxation after PBTI stress, but lower relaxation after NBTI compared to the corresponding b_1 . All these results support the trends schematically shown in Fig. 7.4.

7.5.1 Entire Relaxation

Interestingly, when the ratio b_2/b_1 of each relaxation curve is plotted over the stress time, the resulting curves are ranked according to their electric field during the stress. In Fig. 7.7 (left) equal E_{ox} conditions at various t_{str} values are connected for better visibility and are separated by dotted lines for different electric fields. Different E_{ox} values ranging from NBTI with -8 MV/cm up to PBTI with $+8\text{ MV/cm}$ result in gradually increasing b_2/b_1 , despite some minor deviations for different device thicknesses. Samples stressed with NBTI feature a b_2/b_1 smaller or equal to 1 due to only a small kink or no kink at all, while on the other hand PBTI stress, shows ratios from 1 up to 20.

Hence, the ratio b_2/b_1 gives a measure of the symmetricity of the relaxation curve. The ratio indicates which section of the relaxation transient the original experiment recorded. If $b_2/b_1 < 1$, the experiment probed the second half of the S-shape, i.e. the long-term relaxation, which is usually the case after NBTI. For $b_2/b_1 \approx 1$, the “main” part of the relaxation was monitored and both the initial as well as the late relaxation phase contribute to the total recovery to about the same degree.

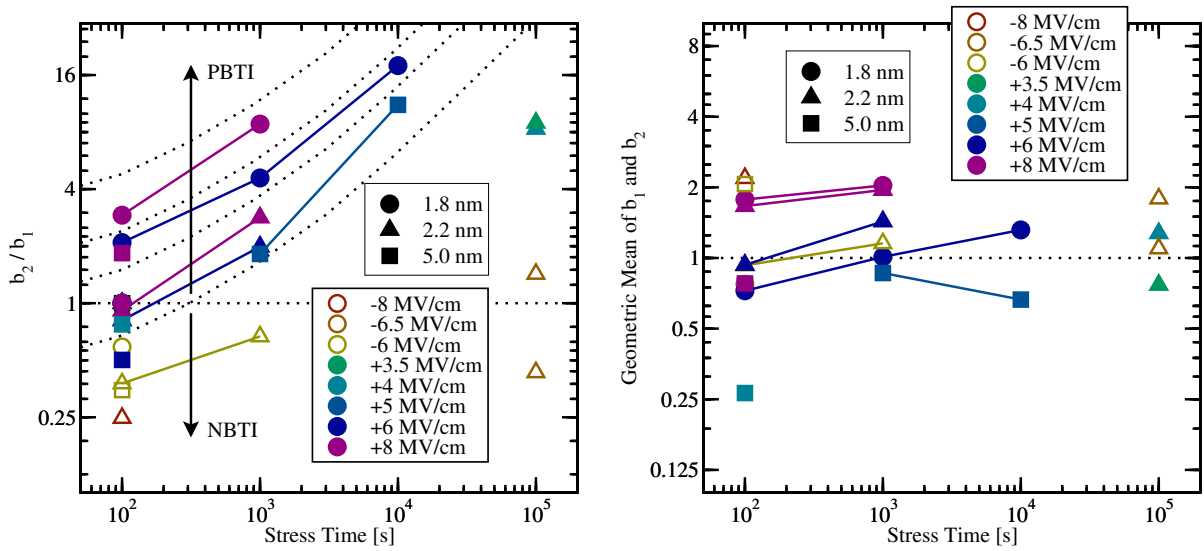


Figure 7.7: **Left:** The ratio b_2/b_1 increases with increasing t_{str} and E_{ox} ranging from NBTI with -8 MV/cm up to PBTI with $+8 \text{ MV/cm}$. NBTI-stressed samples with a negative kink feature a ratio smaller or equal to 1, whereas PBTI-stressed ones possess values from 1 up to 20. Higher ratios are restricted by the maximum allowed electric field E_{ox} before the oxide breaks down. **Right:** Using only a single slope for the recovery characterization obviously eliminates the visible effect of E_{ox} . Hence, the geometric mean of b_1 and b_2 is nearly constant for all analyzed devices despite its weak t_{str} -dependence. This implies that the evaluation of a single slope is not valid for heavier stress conditions because of the asymmetric and limited observation period.

Modeling the recovery with a single slope, which would then be approximately equal to the geometric mean of b_1 and b_2 , clearly obscures the fact that the oxide electric field has an impact not only on the slope, but on the shape of the recovery as well. As depicted in Fig. 7.7 (right), with a mean recovery it is thus only possible to distinguish between the t_{str} . Moreover, the geometric mean requires symmetry of the recovery trace, which is only given under moderate stress conditions.

7.5.2 Change in ΔV_{TH}

Another possibility to evaluate the kink in the recovery characteristics is to determine the slope $d\Delta V_{\text{TH}}/d \log(t_{\text{rel}})$ of the relaxation curve at each point of t_{rel} . This is achieved via linear regression using multiple points of ΔV_{TH} around t_{rel} to obtain the change in its central point $\Delta V_{\text{TH}}(t_{\text{rel}})$. Due to the apparent noise, a multiple-point regression is indispensable; a number of 20, 40, and 80 data points is used for each t_{rel} . Thereby even very small changes in ΔV_{TH} are able to be identified, as illustrated in Fig. 7.8, where the last relaxation curve of a noisy and a less noisy device is depicted. In this figure the linear regression performed with 40 data points around each t_{rel} yields small steps where the slope of ΔV_{TH} suddenly jumps. This issue will be discussed under the aspect of emission times τ_e of certain defects [112] in the next section, where changes of the recovery behavior with varying E_{ox} and t_{str} are due to a change in the emission time rates of the defects [100, 111, 113–115].

Note that using even more than 80 data points around each t_{rel} for the linear regression would even better suppress the noise but on the other hand side would disturb important information at the beginning and at the end of the ΔV_{TH} -curve. Fortunately, the region of interest (around the kink point) lies in the center of a ΔV_{TH} -curve.

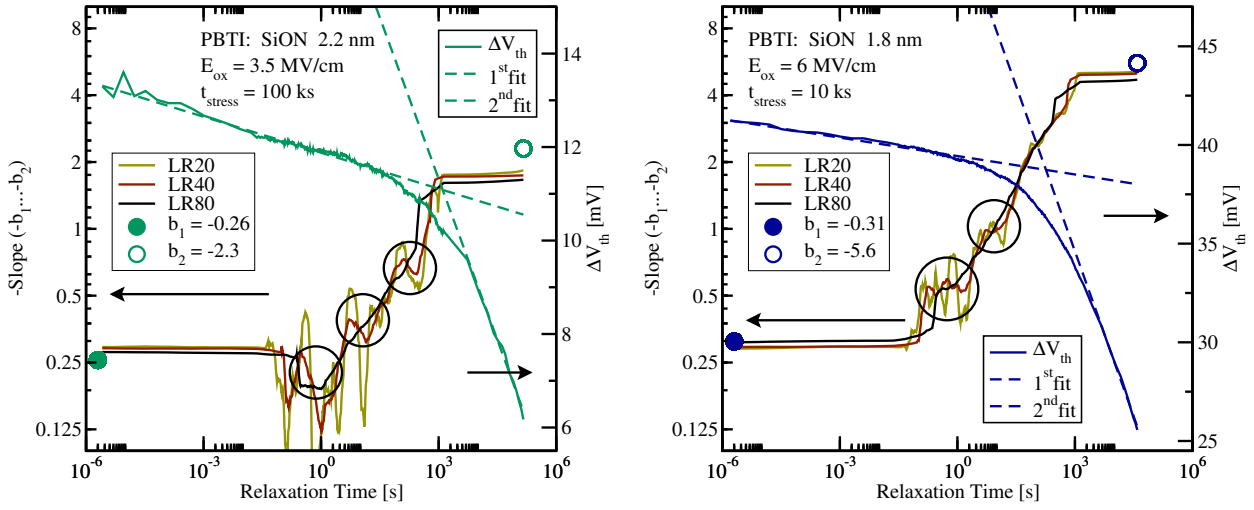


Figure 7.8: The derivative of ΔV_{TH} identifies even very small changes in ΔV_{TH} . To suppress the noise multiple points (20, 40, 80) are used to determine the slope $d\Delta V_{TH}/d \log(t_{rel})$ via linear regression. This is labeled by LR20, LR40, and LR80 above. Using linear regression with more points on the one side smoothes the derivative but on the other hand side removes information at the beginning and at the end of the ΔV_{TH} -curve. This drawback vanishes in the area of interest, around the kink-point. **Left:** For noisy data the slope (LR40) suddenly jumps around $t_{rel} = 1$ s, 12 s, and 170 s, which is marked by circles. Publications dealing with emission time constants of certain defects provide further information [100,111,112]. **Right:** For a thinner device the data is less noisy and the times where the slope changes step-like are more evident, cf. $t_{relax} = 1$ s and 10 s.

7.6 Emission Time Constants

The degraded V_{TH} in small-area transistors with only a few defects relaxes in discrete steps. Each step reveals a hole emission event at the emission time $\tau_{e,i} = \tau_0 \exp(E_{A,i}/k_B T)$ of a particular defect [112, 115]. Larger devices contain a larger number of defects, which in combination with a nearly uniform distribution of the activation energies $E_{A,i}$ yields a log-like recovery behavior as displayed in the top of Fig. 7.9. As there are many different pairs of $\tau_{c,i}$ and $\tau_{e,i}$ within the device, their extraction from the experimental data is discussed first.

By subtracting two recovery traces after stress times $t_{s,i}$ and $t_{s,i+1}$, the fraction of defects with capture time constants with $t_{s,i} < \tau_c < t_{s,i+1}$ is determined first [116], which is shown in Fig. 7.10. By dividing the difference trace into intervals $[t_{r,i}, t_{r,i+1}]$, the fraction of defects having $t_{s,i} < \tau_c < t_{s,i+1}$ and $t_{r,i} < \tau_e < t_{r,i+1}$ is obtained.

To be able to describe the frequency of occurrence of capture time constants τ_c and emission time constants τ_e properly, a large set of long recovery traces with varying t_{str} is needed. The experiments performed cover τ_c from 10^{-6} s up to 10^4 s and τ_e intervals between 10^{-6} s and 10^3 s. This allows for an extraction of the time constants as exemplarily depicted in the bottom of Fig. 7.9.

It is now possible to explain the above mentioned effect with the varying oxide electric field on the basis of Fig. 7.11, where the fraction of ΔV_{TH} due to defects with τ_c and τ_e is plotted as smoothed surface over τ_c and τ_e .

For NBTI with an E_{ox} of -6 MV/cm the surface shows two peaks. One peak covers τ_c and τ_e smaller than $1 \mu s$, while the other more pronounced one clearly illustrates that the largest part

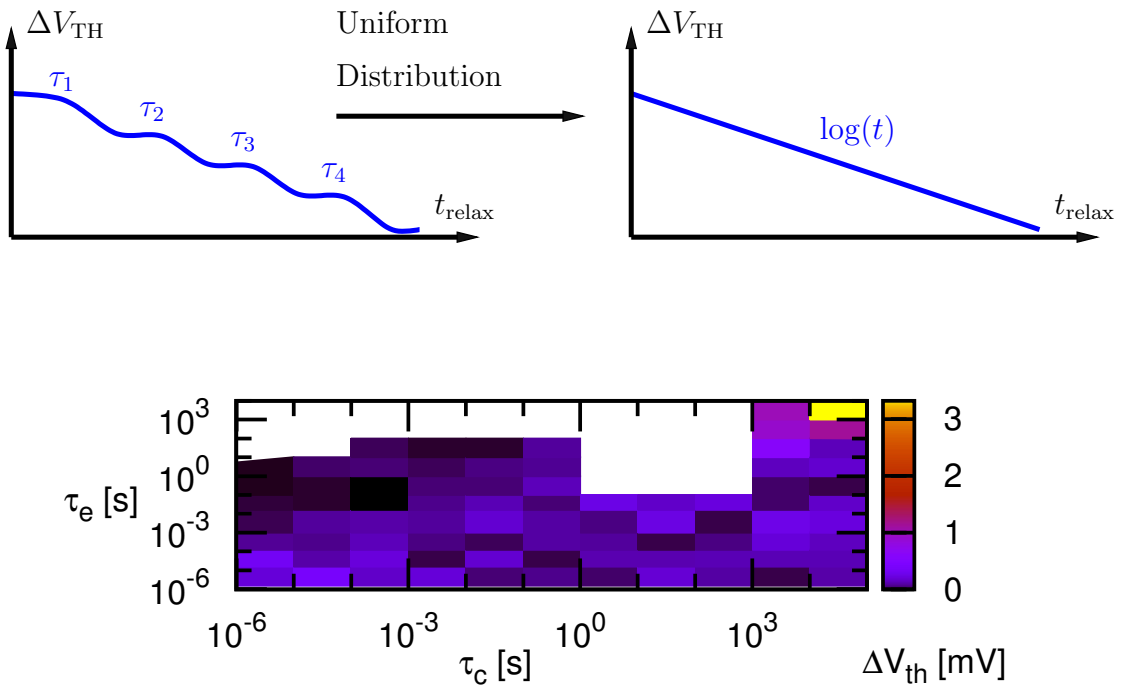


Figure 7.9: **Top:** If there are few defects with emission times τ_i like in small-area transistors [112,115], the relaxation after BTI exhibits discrete jumps. Enlarging the area (more defects) and assuming a uniform distribution of them adds up to a $\log(t)$ behavior, instead. **Bottom:** Map of time constants of capture and emission split into decades of time.

of the degradation was due to defects with τ_c larger than 1 s, which is highlighted by the contour lines below the graph. When comparing the different E_{ox} for PBTI for τ_c covering time constants between 10^2 s and 10^3 s, the peak of 6 MV/cm mainly consists of $\tau_e > 10$ s, while it is widened for 8 MV/cm towards smaller τ_e . This supports the hypothesis of decreased τ_e for higher E_{ox} after PBTI stress, which appears as faster long-term recovery.

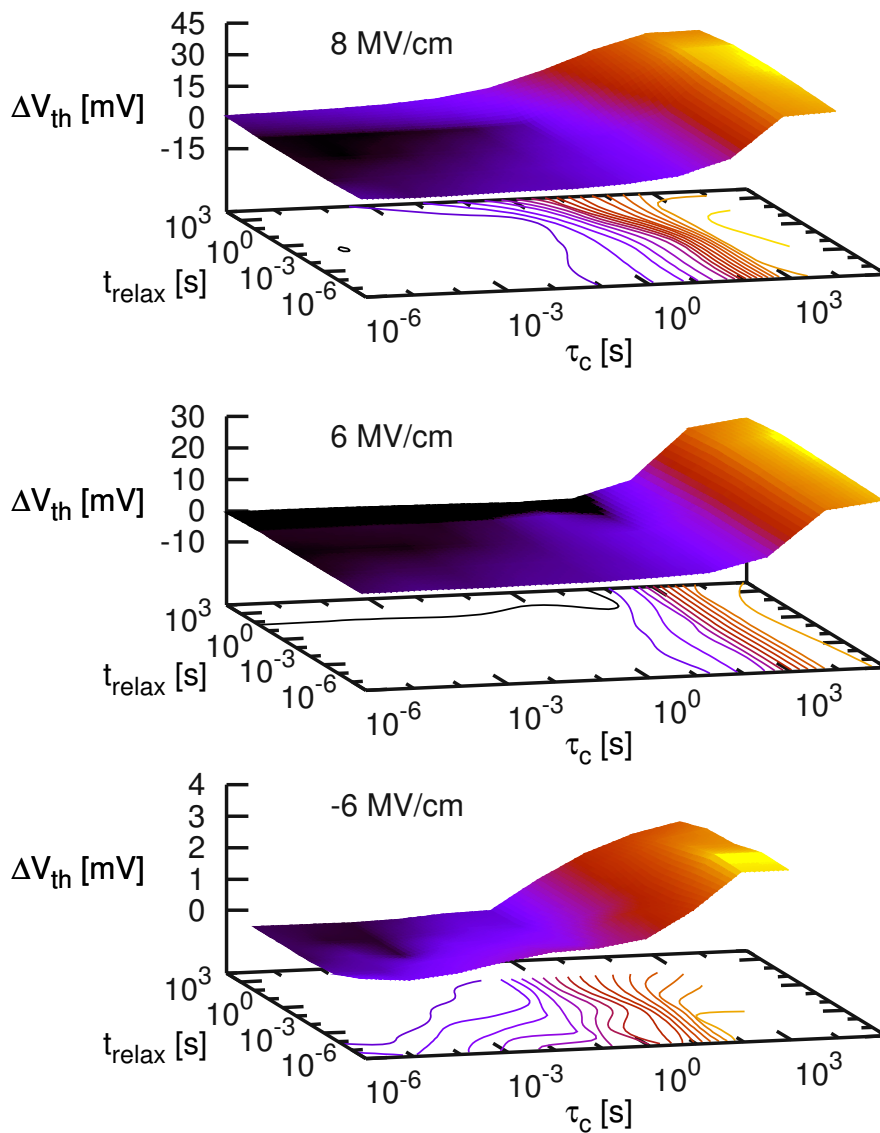


Figure 7.10: Evaluation of the shift between two relaxation curves after stress times $t_{s,i}$ and $t_{s,i+1}$ yields the fraction of defects with capture time constants with $t_{s,i} < \tau_c < t_{s,i+1}$. These ranges of capture time constants of certain defects are depicted as function of t_{rel} . The contour lines below the three graphs emphasize the amount of defects contributing to ΔV_{TH} . For NBTI with an E_{ox} of -6 MV/cm, the characteristics of t_{rel} are not changed with increasing τ_c , despite some shift along the positive ΔV_{TH} -axis. The maximum ΔV_{TH} values for all τ_c -ranges are obtained for small values of t_{rel} . This implies fast relaxation. On the contrary, PBTI (6 MV/cm) yields a larger degradation and additionally moves the characteristics of t_{rel} towards increasing τ_c . For the largest available τ_c , which covers time constants between 10^3 s and 10^4 s, the maximum of ΔV_{TH} is moved away from the minimum t_{rel} . This maximum marks the beginning of the change of emission time constants τ_e depicted in Fig. 7.11 and is even more pronounced for 8 MV/cm.

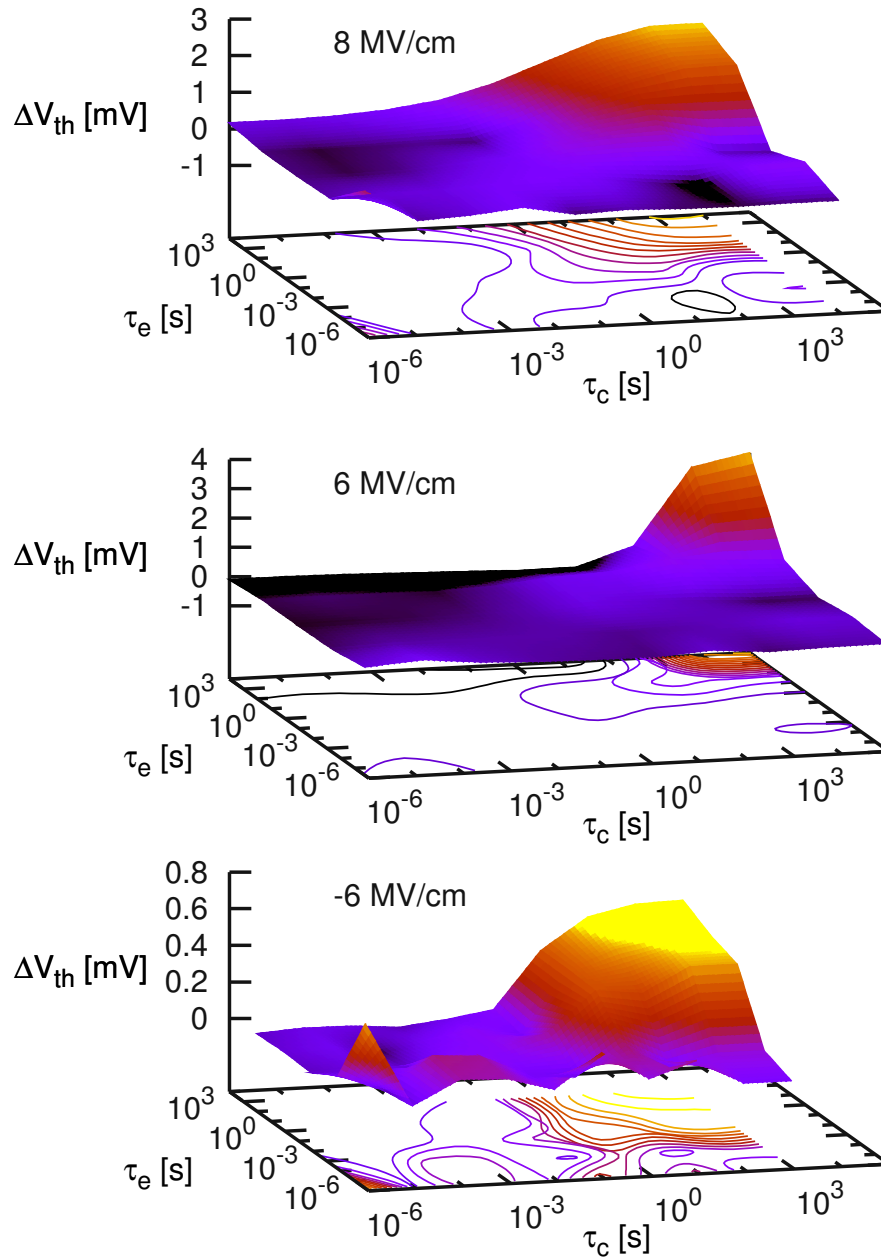


Figure 7.11: The fraction of defects having $t_{s,i} < \tau_c < t_{s,i+1}$ and $t_{r,i} < \tau_e < t_{r,i+1}$ is depicted for three different oxide electric fields. The contour lines below the graphs highlight the biggest changes of ΔV_{TH} . Both surface and contour lines are smoothed for a better visualization. It is shown that the oxide electric field is related to the magnitude of τ_e . Increasing E_{ox} yields a shift of the peak towards smaller τ_e , which corresponds to our monitored increased recovery at larger t_{rel} . Note that only for 6 MV/cm a full set of τ_c and τ_e is available and therefore the map has to be truncated in order to be comparable with the case 8 MV/cm.

7.7 Conclusions

The relaxation behavior of stressed pMOSFETs depends on the oxide electric field and stress time of the performed stress. Especially when dealing with PBTI, the harsher the conditions of stress the later the device starts to relax significantly. By using the limited observation period for NBTI and PBTI as part of the full recovery shape the experimental findings can be explained. The full features of the recovery curve can only be identified after moderate BTI stress, where the relaxation after a certain time accelerates, and slows down again later. Furthermore, deeper analysis of the relaxation characteristics provides information on the distribution of capture and emission times of the defects assumed responsible for BTI. For the case of the PBTI measurements presented in this chapter, especially the distribution of emission time constants depends on the applied oxide electric field during stress. A higher oxide electric field shifts and broadens this distribution. This change in the distribution shows that with a deeper understanding of single capture and emission times it might be possible to reveal the actual origin of the BTI phenomenon.

A method for the detection of the change of a real single defect state, e.g. an electron emission, was already reported by Karwath *et al.* more than 20 years ago. They used the deep level transient spectroscopy (DLTS)¹ to observe the emission times of single isolated defects in small-area MOSFETs by step-like current transients [118]. Such a step-like behavior at the emission time of a defect is also obtained by the time dependent defect spectroscopy (TDDS) [111, 115]. Here small devices are repeatedly stressed (100 times or more) and the averaged relaxation curve is then monitored showing the discharging behavior of the single defects. In order to be able to determine the different capture times of the defects, different stress pulses have to be applied [12]. When a large number of stress and relaxation sequences are collected on a map, the emission times can be obtained. These maps are similar to those presented in this chapter [112, 114, 116]. The major difference lies in the size of the investigated samples. In larger MOSFETs the averaging of the relaxation curves is neither necessary nor reasonable, because already inherent due the large number of defects present there [115]. Although the superposition of many defects is not yet fully understood, it was shown in [116] that the discrete step-like recovery observed in small (narrow) devices is indeed comparable with the nearly continuous recovery behavior obtained for large (wide) devices. The averaging of many small devices also yields a log-like behavior, giving a very strong hint that the underlying mechanism is the same.

¹This technique was originally developed by Lang [117] to characterize the spectrum of traps regarding their energy and concentration.

Chapter 8

Latest Modeling Attempts - Hole Trapping

In Chapter 3 it was tried to explain BTI by using either the diffusion of hydrogen or a dispersive bond breaking mechanism. In both cases interface states are involved. Unfortunately, the theoretical and experimental analysis of the on-the-fly interface traps (OFIT) technique presented in the last chapter revealed that the aberrations leading to the assumption of fast interface state stress and recovery are due to an artifact of the measurement routine. Since the recovery of BTI, especially its short-term behavior, is not explicable with interface states only, hole trapping models have been added [40, 69, 119, 120]. Today the BTI community does still not agree on how holes contribute in detail. The earliest hole modeling attempts date back to the 1950s, where McWhorter used hole trapping to describe $1/f$ -noise at germanium surfaces [121]. More precisely, $1/f$ -noise was considered as oscillations of the trap occupancy of individual defects caused by capture and emission of carriers. McWhorter's attempt is based on the Shockley-Read-Hall (SRH) theory which was originally developed to model the recombination of bulk defects with an energy E_T inside the bandgap [122]. He extended this theory to also model oxide defects, which feature a trap level within the semiconductor bandgap. The local depth of the oxide defect x_T , measured from the interface, enters the model as a tunneling WKB factor $\exp(-x_T/x_0)$, where x_0 acts as scaling factor.

When assuming a defect at E_T capturing a hole from the reservoir in the substrate, e.g. from E_v , the hole does not have to surmount a barrier because of $E_T > E_v$. For the opposite process, namely the hole emission from the defect, the transition probability is reduced by the Boltzmann factor $\exp(-\beta(E_T - E_v))$. However, the application of this approach to a defect level $E_T < E_v$, which can be assumed for oxide defects, makes the above Boltzmann factor larger than unity in the simplest picture. The hole emission barrier rather vanishes in the case of $E_T < E_v$. In turn the corresponding capture process is now affected by an additional Boltzmann factor $\exp(-\beta(E_v - E_T))$ [123]. The hole capture c_p and emission e_p barriers for both kinds of defect, leveling above E_v for the simple SRH and below E_v for the extended SRH, are all depicted in Fig. 8.1 (left).

When an additional oxide electric field F_{ox} is present, the defect level is shifted with respect to E_v . Since the barrier $E_{vT} = E_v - E_T - q_0 x_T F_{ox}$ is linearly dependent on F_{ox} , the defect may now effectively lie below or above E_v , cf. Fig. 8.1 (right). Unfortunately, the McWhorter model was originally developed for $1/f$ -noise in thick oxides and not designed to explain the strong temperature and bias dependence observed during BTI stress in modern devices with oxide thicknesses of only

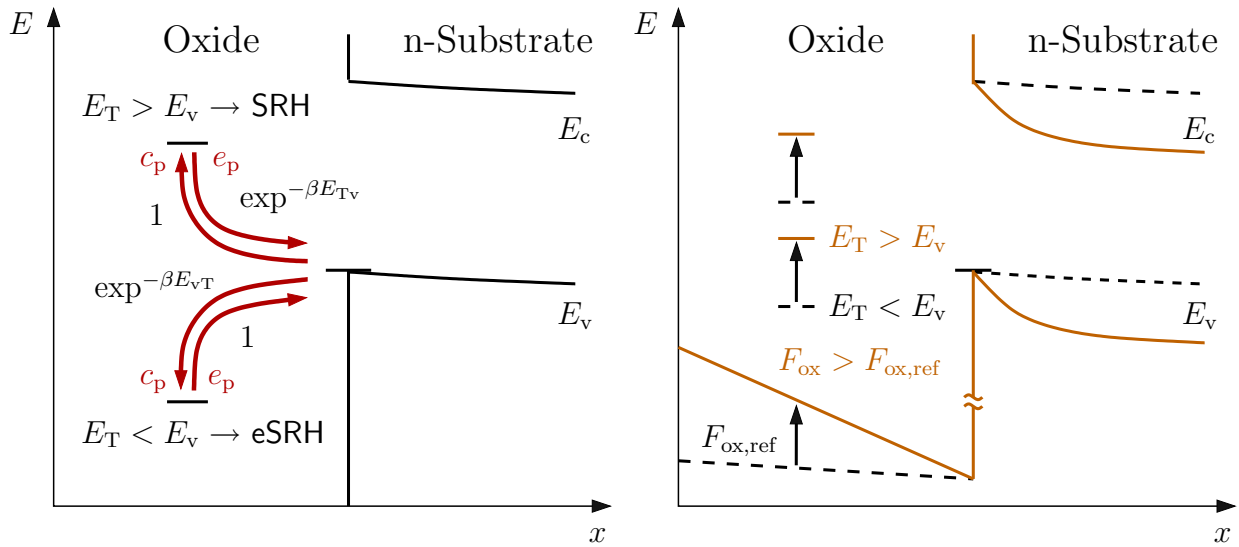


Figure 8.1: Band diagram including a single defect in the oxide. **Left:** Depending on whether the defect level E_T lies above or below the valence band edge E_v , different barriers are obtained. For the characterization of $E_c > E_T > E_v$ the original SRH picture is used, while for defects with $E_T < E_v$ an extension is necessary. **Right:** Applying an additional oxide electric field F_{ox} respective to a reference of $F_{ox,ref}$ shifts the defect level because E_T has to be kept constant within the bandgap. Moreover, the further away from the interface, the more E_T is affected, cf. the oxide bandedges changing from dashed to solid.

a few nanometers, e.g. 2 – 3 nm. In such devices the McWhorter model only gives time constants smaller than a millisecond, which contradicts the measurement results [55].

About thirty years later Kirton and Uren used a modified McWhorter model to explain their random telegraph noise or signal (RTN/RTS) measurements, which characterize the change in the drain current of small-area MOSFETs as a function of time. The times where the signal randomly jumps into the high- and low-current were identified to be Poisson distributed around the expectation value of the capture τ_c and emission τ_e time constants of individual defects respectively. To link this capture and emission kinetics to the observed $1/f$ -spectra, Kirton and Uren proposed the existence of many defects with uniformly distributed time constants on a log scale ranging from milliseconds to days [124]. Since they expected a multi-phonon emission (MPE) process to be responsible for their experimental findings, they added a thermal barrier ΔE_B to the existing SRH model [125–128]. This approach will be continued in the next chapter, where a mathematical description is presented.

8.1 Rate Equations

Based on the existence of oxide defects and the band-to-trap transition possibilities, depicted in Fig. 8.1, already a single defect system has to consider all transitions originating from various band states. This means that the whole conduction or valence band has to be considered, instead of only E_c or E_v . On the basis of the statistical description of the recombination of electrons and holes under the release of energy in terms of lattice vibrations (Shockley-Read-Hall theory [122]), the

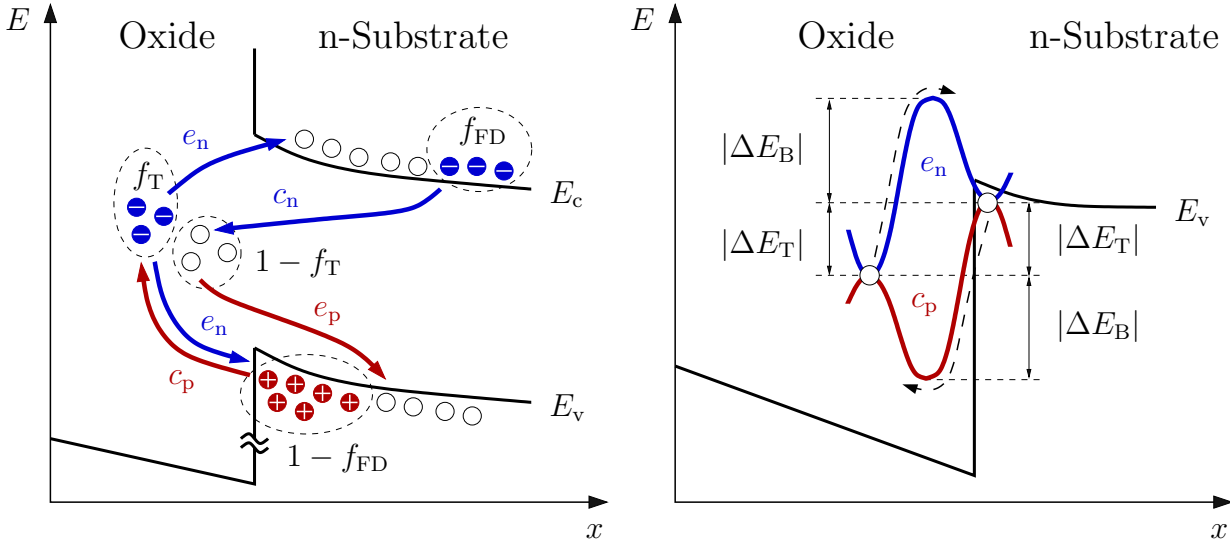


Figure 8.2: **Left:** The rate equations are described on basis of multiple traps in an oxide, charge carriers in an n-substrate, and the corresponding capture and emission coefficients. **Right:** The transition barriers of a hole capture and an electron emission process are equivalent and consist of the trap energy difference according to the state in the substrate ΔE_T and an additional barrier ΔE_B .

determination of effective rates in and out of a specific defect system is possible. The corresponding rate equations are

$$\partial_t f_T = \int_{E_c}^{\infty} [(1 - f_T) f_{FD}(E) c_n(E) - f_T (1 - f_{FD}(E)) e_n(E)] D_c(E) dE \quad (8.1)$$

$$= \int_{-\infty}^{E_v} [(1 - f_T) f_{FD}(E) e_p(E) - f_T (1 - f_{FD}(E)) c_p(E)] D_v(E) dE, \quad (8.2)$$

with the trap occupancy in the oxide f_T and the Fermi-Dirac distribution f_{FD} , which represents the probability of an occupied quantum state in the substrate. Since the Fermi-Dirac distribution is valid in thermal equilibrium and still a very good approximation during BTI, as there is nearly no channel current [129, 130], the distributions write as $f_{FD} = (1 + \exp(\beta(E - E_f)))^{-1}$ and $f_T = (1 + \exp(\beta(E_T - E_f)))^{-1}$. The quantities c_n , e_n , e_p , and c_p stand for the coefficients of electron capture, electron emission, hole emission, and hole capture. The density of states (DOS) is split into a conduction band part D_c and a valence band part D_v .

Assuming detailed balance [122], which means that each process is balanced by its reverse process, both rates have to equal within (8.1) and (8.2). This yields

$$\frac{e_n(E)}{c_n(E)} = \frac{c_p(E)}{e_p(E)} = e^{\beta(E_T - E)}. \quad (8.3)$$

Combining (8.3) with (8.2)¹ and evaluating the integral finally gives the capture time constant $\tau_{c,p}$ of the holes

$$\begin{aligned}\partial_t f_T &= \int_{-\infty}^{E_v} \left((1 - f_T) f_{FD}(E) \frac{c_p(E)}{c_p(E)} - f_T (1 - f_{FD}(E)) \right) c_p(E) D_v(E) dE \\ &= \left((1 - f_T) e^{-\beta(E - E_f)} e^{-\beta(E_T - E)} - f_T \right) \int_{-\infty}^{E_v} (1 - f_{FD}(E)) c_p(E) D_v(E) dE \\ &= \left((1 - f_T) e^{\beta(E_f - E_T)} - f_T \right) \sigma_p v_{th,p} p,\end{aligned}\quad (8.4)$$

with the cross section σ_p and thermal velocity $v_{th,p}$ of the holes with density p . The term outside the brackets can be identified as the capture rate, which can be seen when (8.4) is compared to the simple rate equation of a two-state defect

$$\partial_t f_T = (1 - p_T(t)) k_f - p_T(t) k_r \quad (8.5)$$

with the rate k_f to fill the defect at E_T and k_r for the reverse rate. Furthermore, p_T gives the probability that the defect is actually filled. Consequently, the capture and emission rates can be written as

$$k_{c,p} = 1/\tau_{c,p} = \sigma_p v_{th,p} p \quad (8.6)$$

$$k_{e,p} = 1/\tau_{e,p} = \sigma_p v_{th,p} p e^{\beta(E_f - E_T)} \quad (8.7)$$

or as the relation

$$\frac{1}{\tau_{e,p}} = e^{\beta(E_f - E_T)} \frac{1}{\tau_{c,p}}.$$

In addition to a tunneling coefficient of $\exp(-x_T/x_0)$ to account for the oxide trap depth after [121], the cross section is considered to be thermally activated with a bias independent barrier ΔE_B [124]. Putting these assumptions together yields

$$\sigma_p = \sigma_{p,0} e^{-x_T/x_0} e^{-\beta \Delta E_B}, \quad (8.8)$$

with a constant prefactor $\sigma_{p,0}$ [124, 131]. With the knowledge that whether the defect level lies below or above E_v , different barriers are obtained after Fig. 8.1, equations (8.6) to (8.8) are now used to calculate the capture rates

$$1/\tau_{c,p} = \sigma_{p,0} v_{th,p} p e^{-x_T/x_0} e^{-\beta \Delta E_B} \begin{cases} e^{-\beta(E_v - E_T)} & E_T < E_v \\ 1 & E_T > E_v \end{cases} \quad (8.9)$$

$$1/\tau_{e,p} = \sigma_{p,0} v_{th,p} p e^{-x_T/x_0} e^{-\beta \Delta E_B} \begin{cases} e^{-\beta(E_v - E_f)} & E_T < E_v \\ e^{-\beta(E_T - E_f)} & E_T > E_v. \end{cases} \quad (8.10)$$

As thermal equilibrium is assumed and the density of states is low enough to rule out quantum effects, the Fermi-Dirac-distribution can be replaced by the Maxwell-Boltzmann-distribution [10]

$$p \approx p_{MB} = N_v e^{-\beta(E_f - E_v)} \quad (8.11)$$

¹For this calculation the hole picture is used.

with N_v as effective valence band weight,

$$N_v = 2 \left(\frac{m_{dp} k_B T}{2\pi \hbar^2} \right)^{3/2}.$$

The trapping barrier ΔE_T can further be written as a superposition of the energy distance during flatband $\Delta E_{T,0} = E_{T,0} - E_{v,0}$ and the applied field F_{ox} which changes the relative barrier between semiconductor and oxide, cf. Fig. 8.1 (right) and Fig. 8.2 (right)

$$\Delta E_T(F_{ox}) = E_T(F_{ox}) - E_v(F_{ox}) = \Delta E_{T,0} - q_0 F_{ox} x_T. \quad (8.12)$$

With the help of (8.11) and (8.12) the time constants in (8.9) and (8.10) finally read as

$$1/\tau_{c,p} \approx \sigma_{p,0} v_{th,p} p_{MB} e^{-x_T/x_0} e^{-\beta \Delta E_B} \begin{cases} e^{\beta \Delta E_{T,0}} e^{-\beta q_0 F_{ox} x_T}, & E_T < E_v \\ 1, & E_T > E_v \end{cases} \quad (8.13)$$

$$1/\tau_{e,p} \approx \sigma_{p,0} v_{th,p} N_v e^{-x_T/x_0} e^{-\beta \Delta E_B} \begin{cases} 1, & E_T < E_v \\ e^{-\beta \Delta E_{T,0}} e^{\beta q_0 F_{ox} x_T}. & E_T > E_v \end{cases} \quad (8.14)$$

At first only the part of (8.13) and (8.14), which depends on the relative position of E_T to E_v is discussed. The temperature dependence here is dominated by the thermal barrier $\Delta E_{T,0}$. While the barrier $\Delta E_{T,0}$ determines hole capture when $E_T < E_v$ holds, the barrier $-\Delta E_{T,0}$ contributes to hole emission only when $E_T > E_v$. So the barriers are either relevant for $\tau_{c,p}$ or $\tau_{e,p}$ and do not affect both rates. This is due to the relative position of the energetic defect level and its reservoir, as depicted in Fig. 8.1 (left). When looking at the term $\exp(\pm \beta q_0 F_{ox} x_T)$, it can be seen that the applied field either lowers or rises the barrier, but again the field dependence is only included in either $\tau_{c,p}$ or $\tau_{e,p}$. Additional bias dependencies arise from the surface hole concentration, especially below V_{TH} , and the tunneling coefficient [130].

In a typical BTI stress/relaxation sequence all defects are in thermal equilibrium prior to stress. Due to stress the Fermi level E_f is shifted below E_v . For defects with $E_T < E_v$ the resulting barrier $\Delta E_{T,0}$ can only be balanced by the F_{ox} term in (8.13). After (8.12) this means that energetically deeper defects also need to be located deeper in the oxide in order to become charged during stress, i.e. only defects with $E_T > q_0 \psi_{s, str} - q_0 F_{ox, str} x_T$, where $\psi_{s, str}$ denotes the potential at the interface, are accessible during stress [130]. When the stress is completely removed, E_f is shifted back above E_v and the previously charged defects will be moved back below E_f . According to (8.14) they can be emptied over a small barrier if there is any. Thereby accessible oxide defects now feature $E_T < q_0 \psi_{s, rel} - q_0 F_{ox, rel} x_T$ during relaxation [130]. Thus, the exact defect level is not of particular interest for the capture and emission process. E_T only has to lie inside the accessible energy region, i.e. above E_v for stress and below E_v for relaxation. This means that the conditional part of (8.13) and (8.14) only exhibits a small temperature and field dependence.

It is important to realize that it is the thermal barrier ΔE_B in (8.13) and (8.14) introduced by Kirton and Uren, which gives the required temperature dependence, though this dependence is not fully correct, as will be shown later. To first order, the capture $\tau_{c,p}$ and emission times $\tau_{e,p}$ of the defects are determined by x_T and E_B , making another fact visible: $\tau_{c,p}$ and $\tau_{e,p}$ are correlated, while measurement results determining these times during BTI revealed uncorrelated behavior [112, 116]. This rules out the possibility of describing oxide defects by an extended SRH theory.

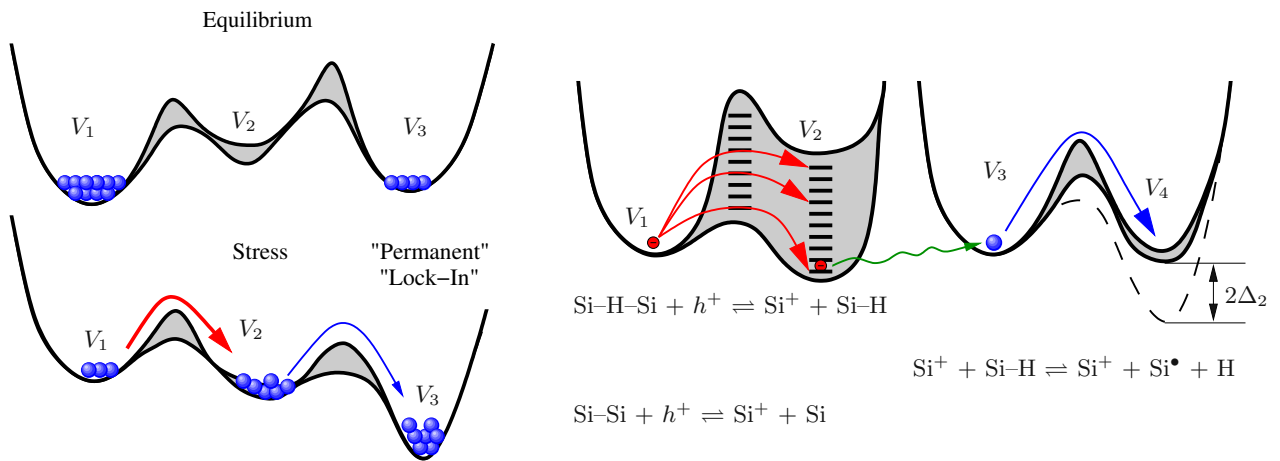


Figure 8.3: **Top Left:** In the triple-well model the second well V_2 is energetically higher than the first and the third well and forms a transitional saddle point. **Bottom Left:** Upon the application of stress V_2 and V_3 are energetically favored and get filled. Since the barrier between V_2 and V_3 is higher than between V_1 and V_2 during relaxation (not shown but comparable to top left), transitions from the second well back to the first well are fast, while the third well represents the permanent component/lock-in. **Right:** In the first step (left double-well) holes are captured at a Si-H-Si bridge or an oxygen vacancy (Si-Si) via a thermally activated process. This captured hole then triggers the release of the hydrogen atom which creates a dangling bond (right double-well).

8.2 Elastic Hole Trapping

So far many hole trapping models have been developed to explain the experimentally observed large time constants and the bias dependence of stress and recovery [40, 69, 119, 120]. Such an approach, developed in the 1990s, is used in the model of Tewksbury which uses elastic hole trapping into preexisting traps [132]. However, the modeling by spatially distributed defect states in the oxide is only feasible in thick oxides, as in ultra-thin oxide layers elastic tunneling gives time constants in the millisecond regime [55] at the most. Moreover, elastic tunneling does not account for the anticipated temperature and bias dependence needed to explain the recent experimental findings [78]. Therefore, a new approach has to be found, i.e. a probably inelastic (thermally activated tunneling) mechanism [98, 133, 134].

8.3 Coupled Double-Well Model

Up to now the appearance of oxide traps and interface states was explained by independent processes adding up as the two components which have been empirically introduced in Chapter 4, i.e. the recoverable and the more or less permanent part of the BTI degradation. Results by Grasser *et al.* indicate that their inducing processes are coupled since their effect cannot be separated by the application of different stress voltages and stress temperatures [134]. By using the basic well-structure of the triple-well model [77], which was already mentioned in Chapter 3.2.1 and is depicted in Fig. 8.3 (left), a new model was introduced consisting of two weakly coupled double-wells [134], cf. Fig. 8.3 (right). During stress holes near the interface can be first trapped to act as a precursor for an interface state. The corresponding reactions are shown in Fig. 8.3 (right) and are based on an oxygen vacancy and a Si-H-Si bridge, respectively [135, 136]. Upon the existence of Si-H

precursors, the second process, i.e. the release of a hydrogen atom, is assumed to be considerably enhanced due to the weaker binding energy of Si–H, compared to that of Si–H–Si². The resulting dangling bonds are poorly recoverable and so account for the demanded permanent component [134].

The major improvement of the coupled double-well lies in the thermally activated hole capture process featuring a dispersive process necessary to explain the wide time scales observed in measurements [11, 134]. Moreover, with this model it was possible to explain a huge amount of experimental stress and relaxation data covering various temperatures, stress voltages, and even device technologies. Unfortunately the physical nature of the coupling inbetween the double-wells remains unclear. This is because a model explaining this coupling requires the consideration of the microscopic behavior of defects.

8.4 Two-Stage Model

For a microscopic model describing the stress during BTI and the relaxation afterwards the oxygen vacancy and the silicon dangling bond are the most likely candidates. The reason for choosing these two defect configurations is that they have been frequently reported to be involved in reliability issues. According to Lenahan silicon dangling bond defects dominate deep levels in the oxide [139]. The oxygen vacancy Si–Si has been used to explain radiation damage [140, 141] and flicker ($1/f$) noise [136, 142] so far.

In the models presented in [124, 136, 143] holes can be captured via a thermally activated multi-phonon emission (MPE) process into states deep in energy but close to the interface, named border traps [99, 144]. Since the MPE process, which will be elaborately explained in Appendix D, was originally derived for bulk semiconductors [125], it cannot be directly used.

Therefore, Grasser *et al.* proposed a two-stage model in [98], which contains an extended MPE process, named multi-phonon field assisted tunneling (MPFAT) process which is similar to the one used in [126, 145]: After [125, 145] the probability of a thermionic transition of a hole over a barrier ΔE_B is $\exp(-\beta\Delta E_B)$. When applying an oxide electric field, the transition probability is further found to be increased by $\exp(F_{\text{ox}}^2/F_{\text{ox,ref}}^2)$, with the oxide electric field F_{ox} and a scaling factor $F_{\text{ox,ref}}$ [98]. This MPFAT process is schematically depicted in the first stage of Fig. 8.4 (state 1 \rightarrow 2). From a defect point of view, the initially assumed neutral oxygen vacancy is charged positively via hole capture. The excess energy of the defect system is subsequently released by structural relaxation [98, 146].

In this way positive E' centers (state 2) are created which can now emit a hole and transfer to (state 3). Being at (state 3) the neutralized defect has either the choice to capture a hole and act as a switching trap by hopping between state 2 and 3 [147], or to fully relax back to its initial precursor state again (state 2 \rightarrow 3 \rightarrow 1). Each path finally leads to (state 1). Therefore, this stage-one describes the recoverable part of the charge trapping.

When the two-stage model is fitted to experimental data of SiO₂, SiON, and high-k devices, the strong (quadratic) voltage and (linear) temperature dependence is predicted correctly supporting the theory of a broad distribution of energetic defects. By involving both oxide charges and interface states contributing to BTI, the model is furthermore able to describe the asymmetric behavior during stress and recovery and the strong bias sensitivity during recovery [98].

²This assumption may hold near the interface. In the bulk, hydrogen is actually weaker bound in Si–H–Si than in Si–H [137, 138].

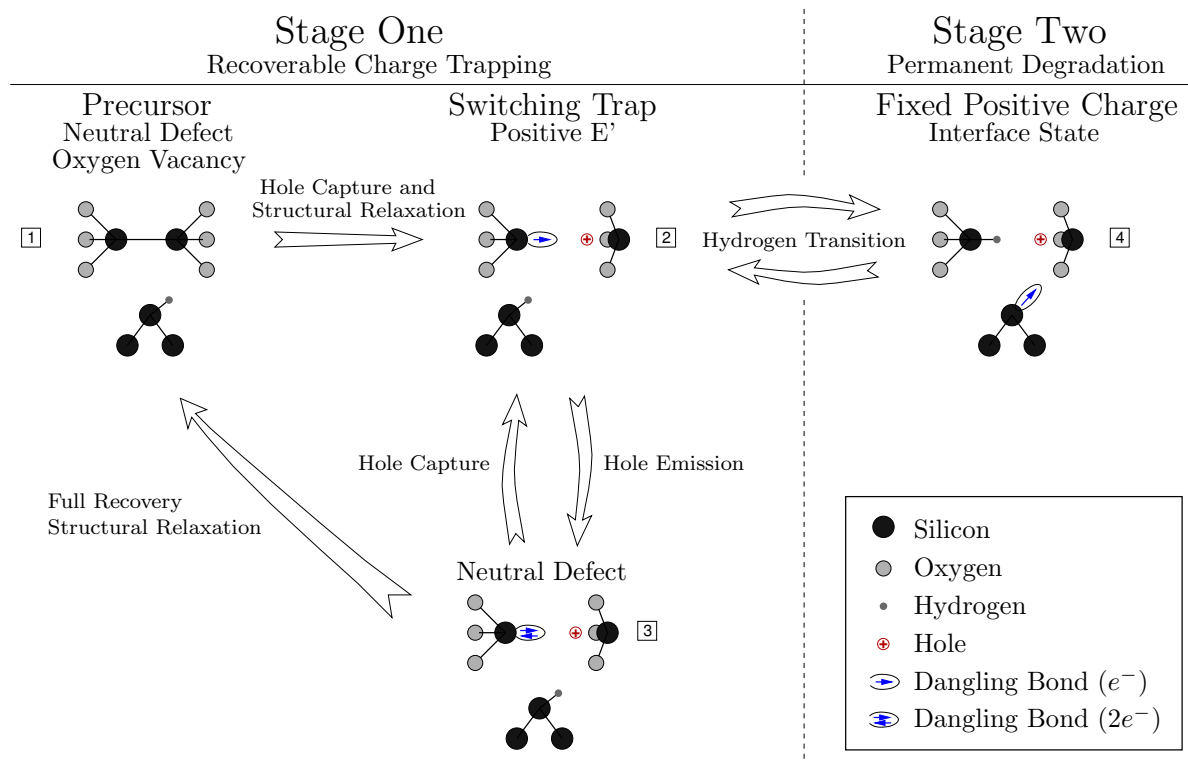


Figure 8.4: The two-stage model starts with a neutral precursor in (state 1). Upon hole capture, the Si-Si bond breaks and a positively charged E' center is created (state 2). Upon hole emission (electron capture) the E' center is neutralized (state 3). Then being in this state, there are two options. Either hole capture again moves the defect back to (state 2), making it act as switching trap, or the defect structure totally relaxes back to its equilibrium configuration (state 1). The transition between stage-one and stage-two is assumed to be via hydrogen transition between state 2 and 4. The dangling bond of the switching trap in (state 2) can capture a hydrogen atom, leaving back a dangling bond at the interface. This passivation by the hydrogen effectively locks the defect state in the positive charge (state 4).

The coupling is established via the transition of a hydrogen atom located at the interface between state 2 and 4. This transition is determined by the hole concentration (positive E' centers) and by the number of hydrogen passivated silicon dangling bonds both available at the interface, i.e. the occupancy of (state 2). When the defect is moved from (state 2 \rightarrow 4), the dangling bond of the oxide defect becomes passivated, leaving back a P_b center [139]. Since a P_b center is a rather stable configuration compared to the switching trap, i.e. the transition rates (state 2 \leftrightarrow 4) are larger than the switching trap rates (state 2 \leftrightarrow 3), the positive defect is hence locked. Consequently, an increased number of defects being in (state 2) favours the creation of permanent states. Mathematically the transition between state 2 and 4 is modeled by thermal activation over a field-dependent barrier, which besides the different ground states E_2 and E_4 , and its corresponding dissociation barrier E_d heavily depends on the applied field F_{ox} , cf. Fig. 8.5 [98].

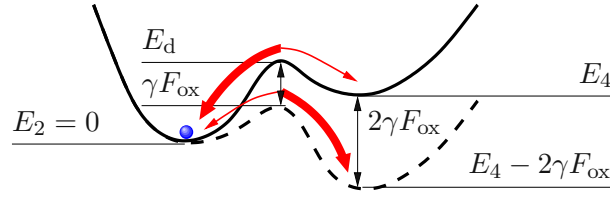


Figure 8.5: The hydrogen transition between state 2 and 4 is modeled by assuming a field-dependent thermal transition over a barrier E_d [98]. Without applied electric field this results in the solid defect configuration. When applying an electric field F_{ox} , the barriers are altered. Consequently the transition rates are changed by the factor $\exp(\pm\beta\gamma F_{ox})$ with respect to the zero-field case. Here ‘+’ holds for the transition ($2 \rightarrow 4$) and ‘-’ for ($4 \rightarrow 2$), with γ being the dipole moment [62]. The transition rates are highlighted by differently thick arrows, indicating that the transition ($2 \rightarrow 4$) is favored, while that of ($4 \rightarrow 2$) is suppressed by the higher F_{ox} .

8.5 Multi-Phonon Emission

Since E' centers upon hole capture may undergo structural relaxation [148], a description of the restructuring process of the concerned defect center is required. By using first-principles density function theory (DFT), Schanovsky *et al.* thoroughly investigated the hole capture process of various point defects embedded in an orthorhombic alpha-quarz supercell structure of 72 atoms [149–151]. During such a hole capture the electronic and vibrational state of the defect system change at the same time, leading to a so-called vibronic transition. This vibronic transition can be modeled using the Born-Oppenheimer approximation and the Franck-Condon (FC) principle [152, 153]: Due to the different masses, electrons and holes only take femto-seconds to switch their states, while defect centers respond with a factor of 100 slower ($1 \cdot 10^{15}$ s versus $1 \cdot 10^{13}$ s). As a consequence, the electrons are able to immediately follow the potential of the defect centers, i.e. they are always in equilibrium compared to the defect centers. On the contrary, the structure of the defect does not change during an electronic transition, which is illustrated by the vertical transition arrow between two different electronic states Fig. 8.6 (left). In this figure the total potential energy of a defect E_{tot} is modeled as a quantum harmonic oscillator featuring the eigenenergy levels of the defect’s vibronic states $E_n = \hbar\omega(n + 1/2)$ with $n = 0, 1, 2, \dots$. According to the Franck-Condon principle, a change of the electronic state, i.e. when moving from one to another harmonic oscillator, at the same time causes a change of the vibrational state and with it a change of the equilibrium position of the defect center [141, 154, 155]. This is known as electron-phonon coupling³. Mathematically, the vibronic transition from the electronic state 1 to state 2 can be derived from Fermi’s golden rule

$$k_{1\alpha \rightarrow 2\beta} = \frac{2\pi}{\hbar} |\langle \eta_{2\beta} \phi_2 | V' | \phi_1 \eta_{1\alpha} \rangle|^2 \delta(E_{2\beta} - E_{1\alpha}), \quad (8.15)$$

where the first index denotes the electronic state and the second index the vibrational state of the electronic $|\phi_a\rangle$ and vibrational $|\eta_{ab}\rangle$ wave functions. After [149, 152], the matrix element of the transition rate in (8.15) can be split into the electronic matrix element represented by a WKB tunneling term and the Franck-Condon overlap factor $|\langle \eta_{2\beta} | \eta_{1\alpha} \rangle|^2$. To consider all possible transitions, the overlap factor has to be calculated for each initial and final state combination, followed by thermally averaging over all initial vibrational states⁴ and then summing over all final vibrational

³In other words, the electron-phonon coupling is the reason why the relative position of two atoms has to change when their bond is altered by removing or adding an electron, to obtain thermal equilibrium again.

⁴The defect system is considered to be in thermal equilibrium before the transition.

states [149,156]. Where the initial and the final total energies are equal [125,153,157], Dirac peaks are obtained whose contour line is called line-shape function (LSF), which describes the broadening of the absorption spectra. Multiplying the WKB term, which approximates the electronic matrix element, and the LSF finally yields the transition rate

$$k_{1\alpha \rightarrow 2\beta} = \underbrace{\frac{2\pi}{\hbar} |\langle \phi_2 | V' | \phi_1 \rangle|^2}_{\text{WKB}} \underbrace{\text{avg}_{\alpha} \sum_{\beta} |\langle \eta_{2\beta} | \eta_{1\alpha} \rangle|^2}_{\text{LSF}} \delta(E_{2\beta} - E_{1\alpha}). \quad (8.16)$$

8.5.1 Approximation of the Vibronic Transition

Basically, the calculation of the LSF via DFT is feasible, but since the motion of a polyatomic structure, especially at $T > 0$, is highly complex to treat, simplifications need to be made. By limiting the movement of the defect system to only one vibrational mode (single-mode coupling), the defect transition can be modeled along its most dominant reaction path or coordinate [151,153]. The total energy E_{tot} as a function of corresponding reaction coordinates (RC) can be further approximated by parabolic potential energy curves (PEC) [158], like schematically depicted in Fig. 8.6. Though originally used for small distortions around the equilibrium, such an harmonic approach is also able to model strong distortions of the defect system [126,159].

The two solid parabolic potentials in the left of Fig. 8.6 are now given by

$$V^0(q) = \frac{1}{2} M \omega_1^2 (q - q_1)^2 + E_{\text{min}}^0 \quad (8.17)$$

$$V^+(q) = \frac{1}{2} M \omega_2^2 (q - q_2)^2 + E_{\text{min}}^+ \quad (8.18)$$

with the mass M and the vibrational frequencies ω_1, ω_2 of the defect system. The minimum of $V^0(q)$ corresponds to the initial defect configuration. When for example examining hole capture, the defect system has to change from $V^0(q_1)$ into its charged configuration $V^+(q)$. This can be achieved by applying a bias which shifts the uncharged defect configuration (solid V^0) with respect to the charged configuration upwards (dashed V^0). When assuming $T = 0$, i.e. there are no phonons, the tunneling process can only occur when the shifted ground state crosses the positive configuration. Starting from $V^+(q_1)$, structural relaxation to the minimum $V^+(q_2)$ takes place. This is accomplished by the emission of phonons. Fowler *et al.* used this picture to model electron tunneling between semiconductor bands and insulator traps at the interface, i.e. band-to-trap tunneling, followed by structural relaxation [160].

8.5.2 Radiative Multi-Phonon Emission

The process described above may be also triggered by an optical excitation, which is usually called multi-phonon emission (MPE), and is illustrated in the center of Fig. 8.6. Following the FC principle, the radiative transition takes place at constant q and moves the defect configuration from the minimum $V_1(q_1)$ up to the V_2 -curve via photon absorption. The necessary energy of this photon can be obtained from the general binding energy $E_B(q) = V_2(q) - V_1(q)$, which is derived in Appendix D.1 as

$$E_B(q) = E_{21} + S \hbar \omega_2 + \frac{1}{2} M \omega_2^2 \left((2q_1 q_2 - q_1^2 - 2q q_2 + q^2) - R^2 (q - q_1)^2 \right). \quad (8.19)$$

Evaluated at q_1 gives an energy of $\epsilon_{12} = E_{21} + S \hbar \omega_2$, with $E_{21} = E_2 - E_1$. This photon energy exceeds the energy needed for a simple electronic SRH transition, where the barrier only results

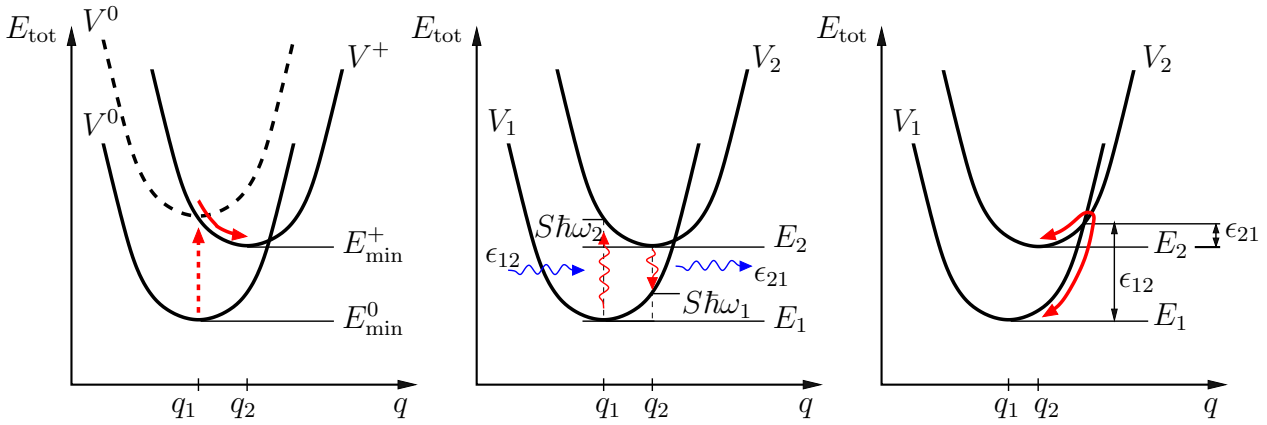


Figure 8.6: The total energy E_{tot} as a function of the reaction coordinate q reveals various transition possibilities of certain defects systems [141, 151, 154, 155, 157]. **Left:** Band-to-trap tunneling is modeled via a two-stage process of tunneling followed by structural relaxation. The dashed line symbolizes the shift of the initial defect system by an applied bias. **Center:** By absorbing or emitting a photon of the energy ϵ_{12} or ϵ_{21} , respectively, the defect state can be changed (multi-phonon emission (MPE)). Subsequent structural relaxation always restores the system to the respective equilibrium in both cases. The emitted energy is called relaxation energy E_R . In the case of linear coupling ($\omega_1 = \omega_2$), $E_R = (\epsilon_{12} - \epsilon_{21})/2$. **Right:** Without optical excitation or emission the same mechanism is called non-radiative multi-phonon (NMP) process. The transition energies ϵ_{12} and ϵ_{21} required have to be supplied by phonons. Classically, the defect has to overcome the barrier determined by the intersection point of the parabolas with reference to E_1 and E_2 .

from the difference of the corresponding energy levels, E_{21} , by $S\hbar\omega_2$. To obtain defect equilibrium within V_2 , $S\hbar\omega_2$ has to be released via structural relaxation (phonons) afterwards. Therefore $S\hbar\omega_2$ is also called relaxation energy. The Huang-Rhys factor S in it gives the number of photons emitted after the FC transition [153] and determines the strength of the electron-phonon coupling [157].

Now the loop can be closed by a photon emission of ϵ_{21} and again structural relaxation ($S\hbar\omega_1$) back to $V_1(q_1)$. Consequently, $\epsilon_{21} = E_{21} - S\hbar\omega_1$. Note that the energy of the two photons ϵ_{12} and ϵ_{21} differs by the sum of the two relaxation energies which are generally not equal due to non-linear electron-phonon coupling.

When electron-phonon coupling is neglected as done in the SRH model, the defect equilibrium does not change. This can be seen by modifying Fig. 8.6 (Center) such that $q_1 = q_2$. Consequently, the photon energies have to be equal now and the relaxation energy $E_R = 0$. As already known, such a harsh approximation it is not able to explain the experimental results of BTI. However, since the calculation of transition barriers with $q_1 \neq q_2$ and $\omega_1 \neq \omega_2$, as would be necessary for a physically more correct model, is quite complex, linear electron-phonon coupling will be used instead⁵. Therein still $q_1 \neq q_2$ holds, but the vibrational frequencies are not allowed to change anymore, i.e. $\omega_1 = \omega_2 = \omega$. Using linear coupling simplifies the model picture a lot. For example, the relaxation energy is now constant too, making the difference of the absorbed and emitted photon exactly $2E_R$ [130, 159].

So far the defect system was treated at $T = 0$ K. At higher temperatures ($T > 0$ K) not only the ground states at the minimum but also higher energies are occupied. For these states the absorbed

⁵Different forms of the coupling have been investigated in [124, 130, 159] and their effect on the bias- and temperature-dependence of the transition probabilities will be discussed later within this chapter.

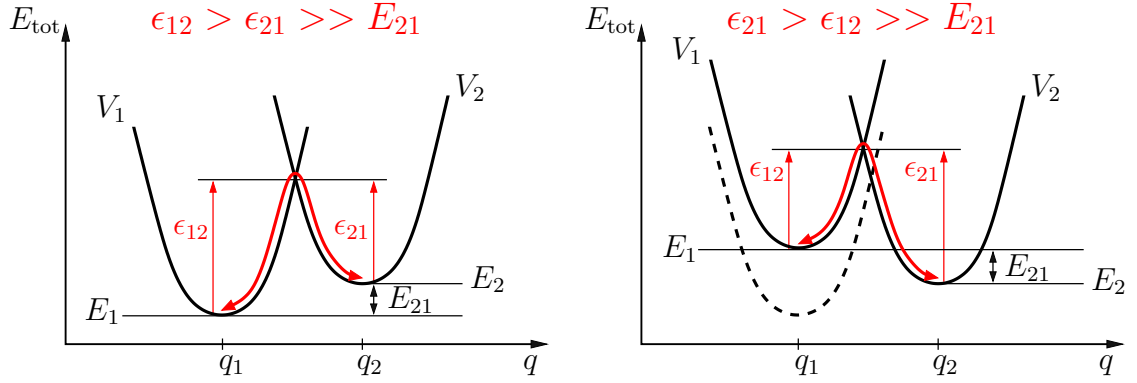


Figure 8.7: The total energy E_{tot} as a function of the reaction coordinate q for non-radiative multi-phonon emission assuming strong ($S\hbar\omega > E_{21}$) and linear coupling ($\omega_1 = \omega_2$). **Left:** Without applied bias $\epsilon_{21} < \epsilon_{12}$ and state 1 is preferred due to its lower total energy. **Right:** Applying a bias induces an oxide electric field which shifts the defect state 1 with respect to state 2. At the same time the intersection point is changed and consequently the barriers ϵ_{12} and ϵ_{21} . For the case shown here the barriers were changed by F_{ox} , i.e. $\epsilon_{21} < \epsilon_{12}$, making the state 2 more likely to be occupied.

and emitted photon energies ϵ_{12} and ϵ_{21} are reduced, which is called thermal broadening of the absorption and emission lines [130].

8.5.3 Non-Radiative Multi-Phonon Theory

An alternative process excludes the absorption and emission of a photon, which is actually the use condition of a MOSFET. This makes it a non-radiative transition (NMP) [112, 130, 161, 162], like depicted in the right of Fig. 8.6. In a classical transition the defect can only surmount the barrier ϵ_{12} or ϵ_{21} between the intersection point of the parabolic potentials and its initial ground state. For linear electron-phonon coupling, i.e. $\omega_1 = \omega_2 = \omega$, these forward and reverse barrier energies are derived in Appendix D.2 to be

$$\epsilon_{12} = \frac{(S\hbar\omega + E_{21})^2}{4S\hbar\omega} \quad (8.20)$$

$$\epsilon_{21} = \frac{(S\hbar\omega - E_{21})^2}{4S\hbar\omega}. \quad (8.21)$$

When shifting the defect level by applying a bias, the defect system in state 1 is shifted with respect to the defect system in state 2. Since the intersection point is changed hereby, the transition rates are directly affected. This approach was already used for the permanent component of the two-stage model depicted in Fig. 8.5 and is schematically shown in Fig. 8.7 for two different bias conditions.

When comparing Fig. 8.6 (right) with Fig. 8.7 (left), a strong difference in $q_2 - q_1 = q_{21}$ can be observed. While for small q_{21} the relaxation energy $S\hbar\omega$ is much smaller compared to E_{21} , it is exactly the opposite for large q_{21} . Depending on which case to deal with, (8.20) and (8.21) can be further approximated. In the first case this yields

$$\epsilon_{12} \approx \frac{E_{21}^2}{4S\hbar\omega} + \frac{E_{21}}{2} \quad (8.22)$$

$$\epsilon_{21} \approx \frac{E_{21}^2}{4S\hbar\omega} - \frac{E_{21}}{2}. \quad (8.23)$$

Since the barriers mainly depend on the difference in electronic energy E_{21} , even quadratically, and not as much on the phonon part contained in the relaxation energy, this case is called weak coupling. Usually one deals with the other case, termed strong coupling, where the barriers are

$$\epsilon_{12} \approx \frac{S\hbar\omega}{4} + \frac{E_{21}}{2} \quad (8.24)$$

$$\epsilon_{21} \approx \frac{S\hbar\omega}{4} - \frac{E_{21}}{2}. \quad (8.25)$$

The barriers here are dominated by the relaxation energy and only linearly depend on E_{21} . This approximation is also visible in Fig. 8.6 (right) for weak and in Fig. 8.7 (left) for strong coupling. When comparing the barriers (8.24) and (8.25) with those within the SRH model (8.13) and (8.14), it can be seen that it is no longer necessary to distinguish whether the trap level is below or above the reservoir level. Furthermore, in the NMP model both barriers of the capture and emission process (ϵ_{12} and ϵ_{21}) depend on the applied field to the same degree with only opposite sign, as can be easily seen in Fig. 8.7. As such the same amount one barrier is lowered is added to the reverse barrier. The resulting field dependence of τ_c and τ_e is hence symmetric for linear coupling. By considering also quadratic electron-phonon coupling terms [112,163], this symmetry is lifted so that one barrier is increased at the expense of the reverse barrier after [130]

$$\epsilon_{12} \approx \frac{S\hbar\omega_1}{(1+R)^2} + \frac{RE_{21}}{1+R} \quad (8.26)$$

$$\epsilon_{21} \approx \frac{S\hbar\omega_1}{(1+R)^2} - \frac{E_{21}}{1+R}. \quad (8.27)$$

with R as ω_1/ω_2 . Unfortunately, this does not solve the undesired correlation of τ_c and τ_e , stated at the end of Chapter 8.1.

8.6 Conclusion

Based on the SRH theory, many models have been developed in order to describe hole capture in oxide defects. Unfortunately, most attempts fail for bias temperature instability, be it because of the missing field dependence or because of the weak temperature dependence. By using a thermally activated barrier E_B in the capture and emission rates of the SRH model, Kirton and Uren already identified that structural relaxation might play an important role to explain the large timescales observed in $1/f$ -noise. Their approach is further developed in the NMP model, where the strong electron-phonon interaction is used to not only explain the temperature dependence of BTI, but also its field dependence. Unfortunately not all effects of BTI can be modeled with the simple rates given by the NMP model. The field dependence experimentally observed exhibits a stronger than linear dependence and furthermore no full decorrelation between the capture and emission time constants is obtained by the simple NMP model. For this reason the NMP model will be extended to so-called metastable defect states in the next chapter.

Chapter 9

Modeling NBTI in High-k SiGe pMOSFETs

In the last chapter it was concluded that the recovery after BTI is the mere consequence of single defects being discharged at certain emission times, which gives a step-like drain current behavior in small devices. The superposition of many of these defects, as observed in large-area devices, yields the typical log-like recovery behavior [11]. The latest attempt to model such defects is based on the non-radiative multi-phonon emission (NMP) theory after [124, 125, 153], cf. Chapter 8.5.3. This theory assumes the conservation of the total energy of a defect or defect system consisting of a strongly coupled electronic and vibronic part [130].

In [112] the NMP model was already shown to successfully reproduce measurement data of small-area devices containing only a few defects. In this chapter it will be shown that the theory also holds for rather complex large-area p-MOSFETs containing a larger number of defects. Such devices have been studied by Franco *et al.* [164, 165] and feature a buried SiGe channel with a thick SiGe quantum well of high Ge-fraction (55 %) and a thin silicon cap below the high-k dielectric in order to reduce NBTI. This type of device is schematically depicted in Fig. 9.1. Devices of this kind were subjected to NBTI stress using various stress voltages and temperatures via the extended measure-stress-measure routine after [18]. For this, a static $I_D(V_G)$ -characteristic is taken first to obtain a reference. After the stress sequences with logarithmically increasing stress times from 2 s to 2000 s the degraded threshold voltage is monitored with a delay of 2 ms.

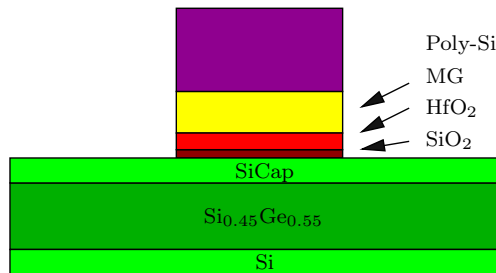


Figure 9.1: Schematic view of the high-k gate-stack device including a thin SiCap and a high-mobility SiGe-layer as quantum well in the channel region to reduce NBTI [164].

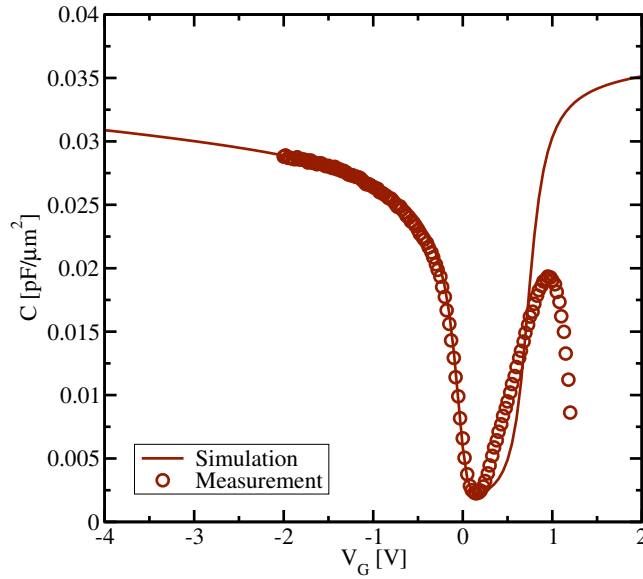


Figure 9.2: The calibration of the pMOSFET simulation model. By using an experimental split- $C(V)$ -characteristic the layer structure can be modeled in 1D. The simulation fits the experimental data very well in the inversion regime, which is required during NBTI. The decrease of the capacitance above 1 V is due to high gate leakage.

9.1 Inverse Modeling

In order to be able to verify the NMP model for the experimental results, inverse modeling is necessary. This is done via $C(V)$ -characteristics of the real device. Unfortunately, the doping profile of the completely processed pMOSFET was not accessible. To incorporate the quite complex NMP model a representative 1D doping profile has to be guessed¹. Since the background doping is small and all the layers below the interface of the pMOSFET are not additionally doped during the fabrication process, diffusion from the source/drain regions towards the SiGe quantum well can take place. Due to many annealing steps the resulting doping becomes complicated which makes the calibration of the simulated layer structure with the proper layer thicknesses and their corresponding dopings an extremely challenging task.

Due to the fact that only a very limited amount of mobility data for SiGe-layers was published so far the mobilities are only roughly approximated and are furthermore considered as constant within the single layers. Based on the measured mobilities of Si and Ge in [89, 166–168], the values for the layers were linearly interpolated for $\text{Si}_{0.45}\text{Ge}_{0.55}$. The used values of μ_i are given in Tab. 9.1 for all layers at a temperature of 400 K.

Despite this approximation the finally obtained $C(V)$ -characteristics of the 1D device catches the trend of the measurement, cf. Fig. 9.2, and fits the experimental split- $C(V)$ very well in the inversion regime. The decrease of the capacitance above $V_G = 1$ V is due to dominant gate leakage. Below $V_G = -2$ V the gate dielectric starts to break down due to the low effective oxide thickness of about 1 nm. Since the NBTI stress conditions dealt with in this chapter are well within this regime, the deviations outside this regime are assumed to be unimportant.

¹The 1D-approximation is valid since V_D is small and the channel is assumed homogenous along the length and width of the device.

Layer	Doping N_A/cm^{-3}	Mobility $\mu_i/\text{cm}^2\text{V}^{-1}\text{s}^{-1}$
SiCap	$5 \cdot 10^{16}$	233
Si _{0.45} Ge _{0.55}	$1 \cdot 10^{17}$	485
Si	$3 \cdot 10^{17}$	203
Si	$2 \cdot 10^{17}$	215

Table 9.1: Details of the used dopings inside the single layers. Based on available measurement data the hole drift mobility was approximated.

9.2 Multi-State Defect Model

Although the non-radiative multi-phonon (NMP) model of Chapter 8.5.3 appears to be the best modeling approach so far, it still suffers from some limitation when used for BTI. First it is not able to fully explain the uncorrelated behavior of the capture and emission times as observed experimentally and second it does not give a stronger than linear field dependence of the capture time constants of the defects though these are observed [169].

When modeling random telegraph noise (RTN) twenty years ago Uren *et al.* [170] suggested that individual states can exist in more than one charge-equivalent, so-called metastable states. Based on this idea the two possible states of the NMP model are now extended to a four-state defect system [112]. Such a multi-state defect model is depicted in Fig. 9.3 (top) for the oxygen vacancy. It contains two metastable defect states $1'$ and $2'$ which belong to the already used stable states 1 and 2, respectively. As can be seen the transitions between 1 and 2 now have to proceed over one of the metastable states. This picture is similar to that already used in the two-stage model, cf. Chapter 8.4, where the oxygen vacancy upon hole capture (state 3) was regarded to be in a kind of metastable state with the choice to either structurally relax or to recapture a hole again. In the multi-state defect model, again the oxygen vacancy is used as switching oxide trap [112, 147].

With the help of a schematic reaction coordinate diagram in Fig. 9.3 (bottom) the transitions of a single defect state are now explained. The neutral defect state 1 and the charged defect state 2 are depicted together with their corresponding metastable states $1'$ and $2'$. Upon the application of a stress bias, the charge transfer reaction from state 1 to $2'$ is favored. This is indicated by the dashed upwards shift of the parabola in Fig. 9.3 (bottom) and leads to a strong electric field dependence of the barrier [112, 157]. As $2'$ is metastable it can relax into its stable form 2 afterwards.

By performing DFT calculations in crystalline SiO₂, Schanovsky *et al.* concluded that the oxygen vacancy does not fulfill all requirements of the multi-state model. This is because its thermodynamic energy level of around 1 eV above the SiO₂ valence band results in a very high barrier for the capture process, which can only be surmounted at very high oxide electric fields. Unfortunately the necessary fields are around 20 MV/cm when assuming the defect to be localized 1 nm inside the oxide [151]. Another problem of the oxygen vacancy is that its neutralized puckered state $1'$ is too unstable to allow a switching trap behavior between state $1'$ and 2, as the defect would rather relax back to its initial state 1 immediately [138].

For the sake of completeness also the hydrogen bridge is briefly discussed. According to first principle calculations its energy level was determined to lie within the Si bandgap, meaning that the defect configuration is already positively charged prior to stress. Therefore the hydrogen bridge is ruled out as possible defect state when dealing with BTI as well [151].

However, no matter what exact defect configuration is responsible for BTI, the multi-state defect model captures the essence of BTI very accurately and will therefore be used in the following. Its

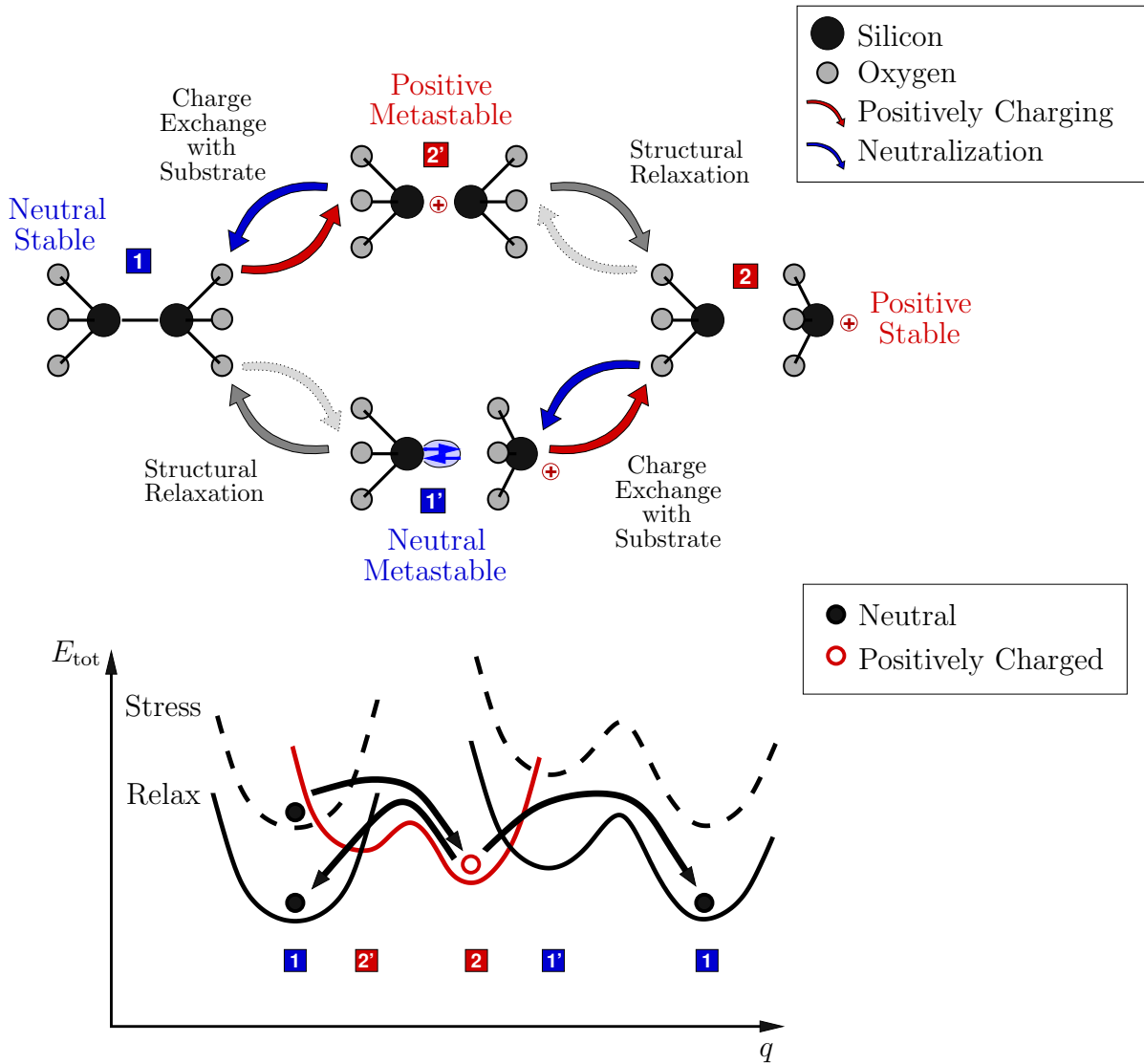


Figure 9.3: **Top:** The improved non-radiative multi-phonon (NMP) model includes a metastable defect state for both charge states. The possible transitions exhibit a charge exchange and/or a structural relaxation. **Bottom:** Schematic reaction coordinate diagram model for a single defect. The different configurational potentials for each defect define the NMP process. The varying bias conditions (stress/relaxation) further change the relative position of the potentials and determine the transition rates from 1 to 2 (over the metastable 1' and 2') and back.

hole capture and emission rates are derived similarly to Chapter 8.1, with the barriers based on the NMP formalism, cf. Appendix D.2

$$k_{12'} = \sigma_p v_{th,p} p \exp^{-\beta \epsilon_{12'}}, \quad k_{2'1} = \sigma_p v_{th,p} N_v \exp^{-\beta \epsilon_{2'1}}, \quad (9.1)$$

$$k_{1'2} = \sigma_p v_{th,p} p \exp^{-\beta \epsilon_{1'2}}, \quad k_{21'} = \sigma_p v_{th,p} N_v \exp^{-\beta \epsilon_{21'}}. \quad (9.2)$$

The indices of the barriers ϵ_{ij} in the rates k_{ij} hold for the corresponding transitions from i to j with the cross section σ_p and the thermal velocity of holes $v_{th,p}$. Furthermore, p and N_v denote

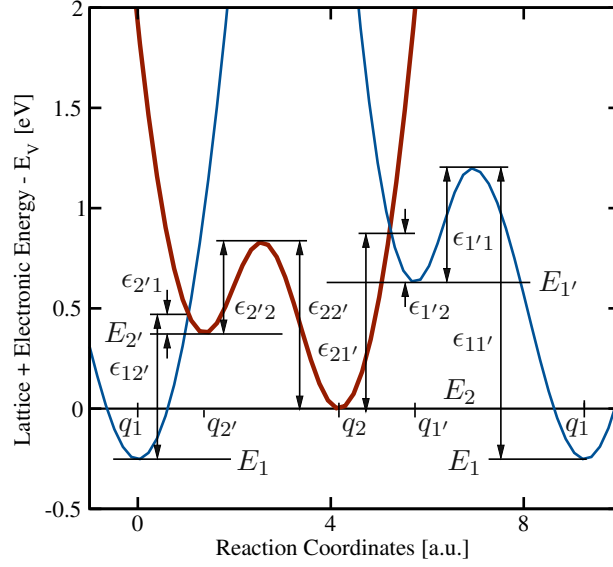


Figure 9.4: The quantities used in the multi-state defect model, taken from [112]. Since the reaction coordinates describing the transition $1 \leftrightarrow 2'$ differ from those describing $2 \leftrightarrow 1'$, the harmonic potentials describing states 1 and $1'$ (blue) are plotted twice.

the hole concentration and the effective valence band weight, respectively. Since the stable and their corresponding metastable states are only separated by a thermal barrier $\epsilon_{ii'}$, these barriers are bias independent. Therefore $\epsilon_{ii'}$ is not calculated via the intersection of the parabolas, as it needs to be done for (9.1) and (9.2), but is an explicit parameter together with an attempt frequency $\sigma_0 \approx 1 \cdot 10^{13} \text{ s}^{-1}$ after [112]

$$k_{1'1} = \sigma_0 \exp^{-\beta \epsilon_{1'1}}, \quad k_{11'} = \sigma_0 \exp^{-\beta(\epsilon_{1'1} + E_{1'} - E_1)}, \quad (9.3)$$

$$k_{2'2} = \sigma_0 \exp^{-\beta \epsilon_{2'2}}, \quad k_{22'} = \sigma_0 \exp^{-\beta(\epsilon_{2'2} + E_{2'} - E_2)}. \quad (9.4)$$

The rates correspond to the barriers and energies in Fig. 9.4. When the effective transition from state 1 to 2 over the metastable state $2'$ is considered this exhibits a two step process, which is proportional to the product of the first rate $k_{12'}$ times the second rate $k_{2'2}$ which has to be divided by the sum of all rates contributing to state $2'$, $k_{12'} + k_{2'1} + k_{2'2} + k_{22'}$ [130]. Neglecting the last rate $k_{22'}$ because it will be smaller than $k_{2'2}$ due to the fixed thermal barrier at all times and adding the path over the metastable state $1'$ in the same manner yields

$$k_{12,\text{eff}} = \frac{k_{12'}k_{2'2}}{k_{12'} + k_{2'2} + k_{2'1}} + \frac{k_{11'}k_{1'2}}{k_{11'} + k_{1'2} + k_{1'1}}. \quad (9.5)$$

Analogously the effective transition from 2 to 1 over the metastable states can be derived to be

$$k_{21,\text{eff}} = \frac{k_{22'}k_{2'1}}{k_{22'} + k_{2'1} + k_{2'2}} + \frac{k_{21'}k_{1'1}}{k_{21'} + k_{1'1} + k_{1'2}} \quad (9.6)$$

These reaction rates then define the specific capture and emission time constants $\tau_c = 1/k_{12,\text{eff}}$ and $\tau_e = 1/k_{21,\text{eff}}$, respectively, within which the single defect is charged and discharged on average.

During stress the effective forward rate $k_{12,\text{eff}}$ can be approximated to the transition over state $2'$, as already indicated in Fig. 9.3 (bottom). However, during relaxation transitions over both metastable states contribute, also indicated by arrows. After Grasser *et al.* [112] switching trap behavior in the multi-state defect model is only observed when both barriers between state 2 and $1'$ are rather small compared to $\epsilon_{22'}$. Consequently, switching traps favor the backward process over $1'$. For defects featuring a large barrier $\epsilon_{21'}$ on the other hand no such switching trap behavior can be observed, since they practically never reach state $1'$, they have to recover via $2'$.

9.2.1 Distribution of Defects

Compared to the very small devices described in [112,116], where a discussion of a single defect makes sense and one is indeed able to resolve the state in which the defect currently is, this is not the case for the large area devices considered in this chapter, where a large number of differently behaving defects is located in the oxide. Therefore, a distribution of defects regarding their characteristics has to be assumed, with all the defects featuring different energies E_1 , $E_{1'}$ and $\epsilon_{1'1}$. Furthermore, different positions x_T inside the oxide and relaxation energies $S\hbar\omega_1$ and $S\hbar\omega_2$ are assumed.

As a reasonable discretization of the defect characteristics is not feasible, the defect properties are considered using a statistical approach, which is shown in Fig. 9.5. Whereas E_1 , $E_{1'}$, and $\epsilon_{1'1}$ are distributed gaussian, x_T , $S\hbar\omega_1$, and $S\hbar\omega_2$ are distributed uniformly within the depicted ranges. All other parameters of the model in (9.1) to (9.4) are taken to be constant.

By choosing 1000 different defects a gaussian profile can already be well approximated, cf. Fig. 9.5 (top left). Further increasing the number of defects (top right) to 100000 yields a nearly perfect gaussian profile. The same numerical improvement is observed for the relaxation energies in Fig. 9.5 (bottom right). However, as will be shown later, a number of 1000 representative defects in a defect band, like depicted in Fig. 9.5 (bottom left), is sufficient to describe the measurement results properly.

When applying NBTI stress, the defect band is shifted with respect to the valence band. Hence only a part of all present defects is able to contribute to the degradation depending on the applied bias conditions. As an example of which fraction of the defects can become charged, the occupied defects are marked as filled within the defect band in Fig. 9.6 (top left) after 2000s of stress. When switching back to relaxation the defect band is lowered again and the defects can become discharged.

However, the defect occupancy is not only determined by E_1 and x_T anymore as it is the case after the SRH-like approach of Kirton *et al.* [124], where all defects up to a certain distance x_T become charged as a direct consequence of the tunneling front due to the WKB factor. The addition of a barrier ΔE_B for all defects only changes the level above and below which the defects can become charged and discharged, respectively. In contrast to this there are also empty defects inbetween the filled ones in terms of both E_1 and x_T after the multi-state defect model, cf. Fig. 9.6 (top left). Unfortunately this means that it is no longer possible to estimate the time dependence of the total degradation as an analytic expression, as it was possible for the SRH approach on basis of x_T in first order. Due to the superposition of many defects, which are further distributed in different quantities, no kind of degradation estimation does make sense for the large devices investigated in this chapter.

Before further discussing the repeated charging and discharging of the switching traps and their time dependence within the performed stress/relaxation cycles, the reservoir conditions of the substrate providing the holes have to be set.

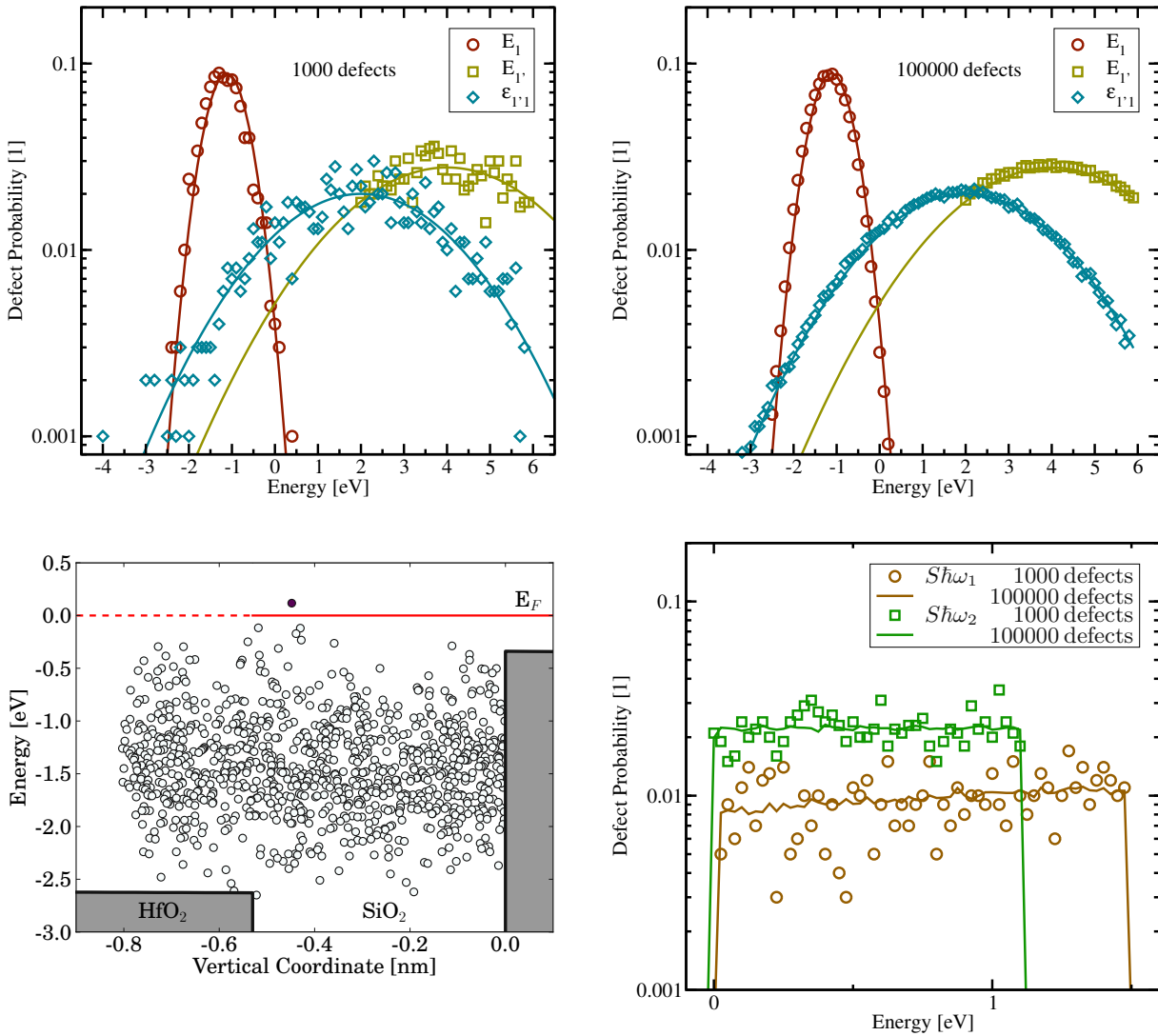


Figure 9.5: Distributed quantities of the defects. **Top Left:** The determining energies and barriers of state 1 are assumed to be gaussian distributed for 1000 defects. Note that the gaussian is intentionally cut off above 6 eV, since barriers above can be not surmounted within the monitored timescale. **Top Right:** When the number of defects is increased by two decades a nearly perfect gaussian distribution is obtained. **Bottom Left:** The defects are uniformly distributed within around 0.8 nm from the interface. Note the already charged defect above the Fermi level for the depicted equilibrium condition. **Bottom Right:** The relaxation energies for both stable defect states $Sh\omega_1$ and $Sh\omega_2$ are uniformly distributed within 0 eV and 1.1 eV or 1.5 eV, respectively.

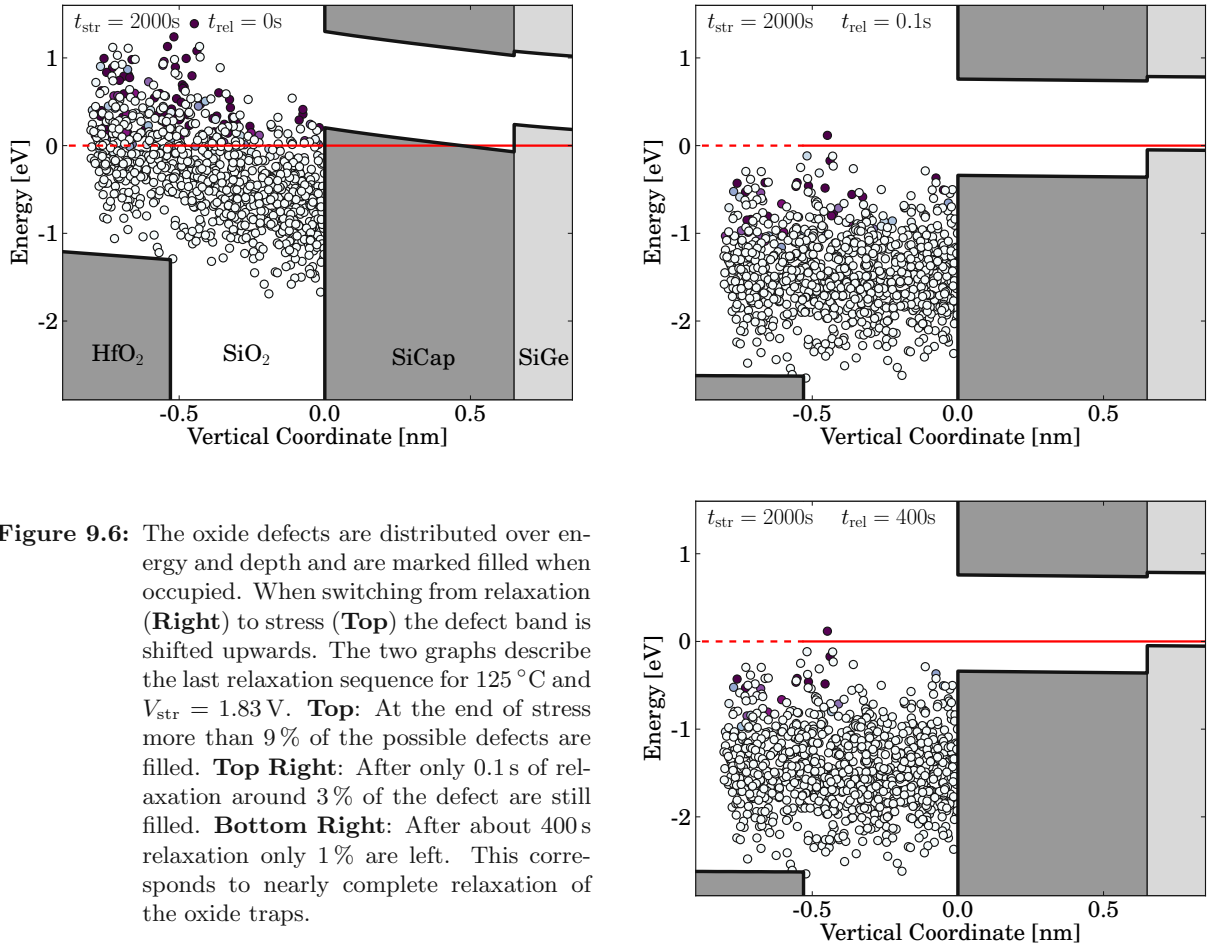


Figure 9.6: The oxide defects are distributed over energy and depth and are marked filled when occupied. When switching from relaxation (**Right**) to stress (**Top**) the defect band is shifted upwards. The two graphs describe the last relaxation sequence for 125 °C and $V_{str} = 1.83$ V. **Top:** At the end of stress more than 9% of the possible defects are filled. **Top Right:** After only 0.1 s of relaxation around 3% of the defect are still filled. **Bottom Right:** After about 400 s relaxation only 1% are left. This corresponds to nearly complete relaxation of the oxide traps.

9.2.2 Reservoir of Holes - Classical vs. Quantum Mechanical Description

Previously [112], the holes were assumed to be energetically located at the valence band edge of the substrate with the defects being filled corresponding to this energy level. Unfortunately this approximation is questionable for a more complex structure like pMOSFETs with a high-k dielectric layer and a SiGe-layer inside the silicon substrate. This necessitates the incorporation of quantum mechanical (QM) confinement. Instead of assuming all holes to be fixed at E_v , it is now distinguished between the contributing subbands, i.e. their different eigenenergies and hole occupancies are considered. To obtain the wave functions of the subbands in the channel of the MOSFET, the Schrödinger and Poisson equation were solved self-consistently using the Vienna Schrödinger Poisson solver (VSP2) [171]. The carrier concentration is calculated by treating the quasi-bound states as a two-dimensional electron gas in equilibrium and the continuum states as a 3D electron gas. The three X valley sorts of the conduction band as well as the heavy hole, light hole, and split off band are taken into account.

In the left of Fig. 9.7 the first five subbands are displayed based on their corresponding eigenenergies of which four are localized in the SiGe-layer. In the right of Fig. 9.7 the first two subbands are depicted for two stress and relaxation conditions. When switching from relaxation to stress the maxima of the subband wave functions move towards the interface, which raises the “hole concen-

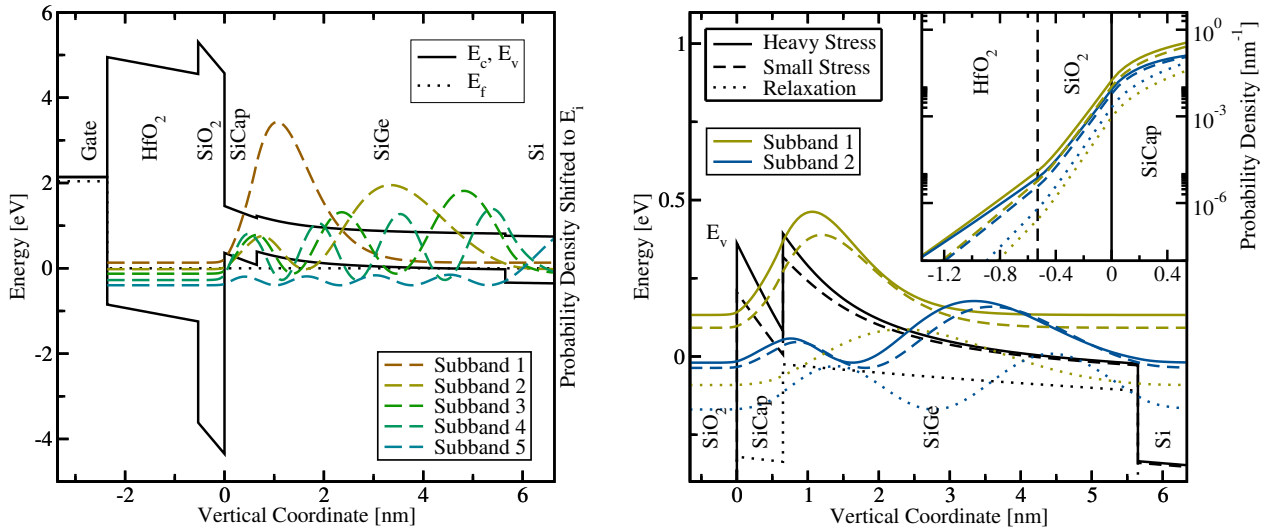


Figure 9.7: **Left:** The NMP-rates are based on the different subbands serving as discrete provider of holes (shown for stress case). Only the first five subbands are displayed based on their corresponding eigenenergies of which four are localized in the SiGe-layer in the substrate. **Right:** When switching from relaxation to stress the maxima of the subbands move towards the interface, increasing the hole concentration. The inset shows the penetration of the wave function on a log-scale. The kink of the wave function in the oxide is caused by the layer structure.

tration” in the oxide. The penetration of the wave functions is plotted on a log-scale to show the transmission probability. It can be clearly seen that the contribution of the subbands decreases with increasing order. As the bandbending inside the oxide due to the charged defects can be neglected, the WKB approximation is valid here and closely matches the transition probability of the wave function. Due to the lower computational efforts, the WKB approach is therefore used in the following.

In order to calculate the occupancy of the entire defect band during a certain bias condition in time, the effective rates in (9.5) and (9.6) have to be evaluated for each subband for each individual defect. The effective rates of a single defect have then to be summed up over all subbands and determine the single defect occupancy. The occupancy of the entire defect band finally gives the observable degradation.

9.3 Results

To validate the multi-state defect model described in the last section, experimental data with logarithmically increasing stress times for different temperatures T and stress voltages V_{str} is essential. This measurement data is then fitted by using a single parameter set. Eventually, the calibrated parameter set is shown in the following to successfully account for all performed measurements.

For a given temperature and voltage, as illustrated in Fig. 9.8 (left), it is possible to simulate a complex eMSM-sequence, consisting of logarithmically increasing stress times, with very good agreement. Furthermore, the last relaxation sequence after more than 2000 s of stress is compared at different temperatures and for different stress voltages, cf. Fig. 9.8 (right). Despite device to device deviations, as various MOSFETs have to be used for the measurements to avoid pre-stress,

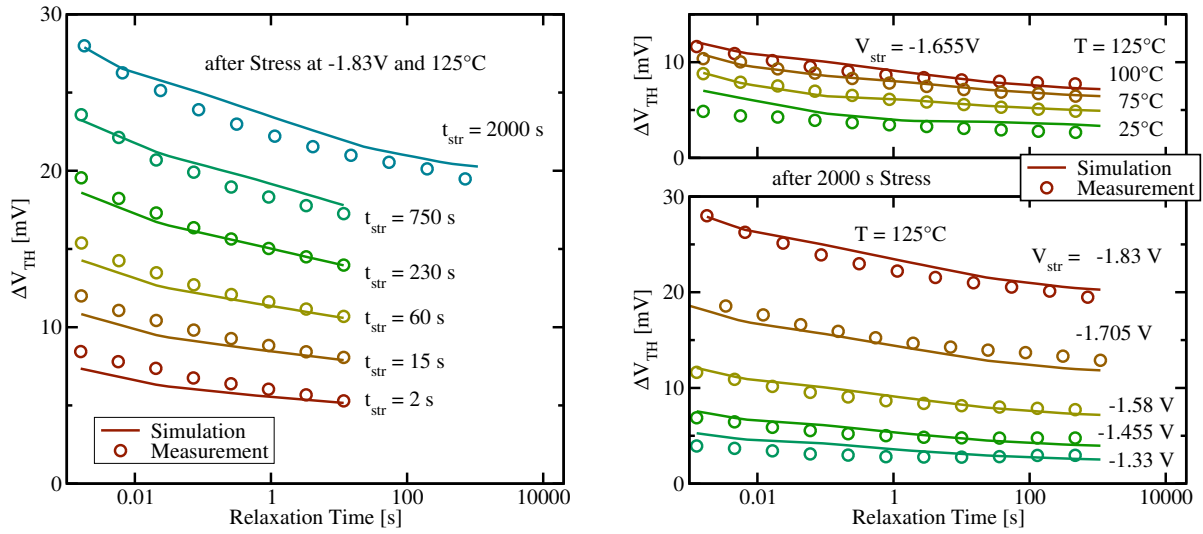


Figure 9.8: **Left:** The relaxation sequences of an measurement-sequence with logarithmically increasing stress times are plotted versus the simulation results. The same calibrated parameter set of Chapter 9.2.1 is applied for all simulations, yielding sound agreement. **Right:** The last relaxation sequence is depicted for different temperatures (top right) and stress voltages (bottom right). It can be clearly seen that the measurement is very well reproduced using the multi-state defect model.

it can be clearly seen that the measurements are very well reproduced using the multi-state defect model in combination with a broad defect distribution.

It was already mentioned that an investigation of the time dependence of a single defect is not reasonable in large device. However, the relaxation behavior observed in Fig. 9.8 can be modeled by the superposition of defects with different emission time constants. This is confirmed by Fig. 9.6, where the percentage of the charged oxide defects is correlated with the last relaxation sequence for 125°C and $V_{str} = 1.83$ V. At the beginning of the relaxation 9% of the possible defects contribute, while after about 400s relaxation only 1% are left, which is equivalent to nearly complete relaxation of the switching traps. The permanent part left can be explained by recalling the two-stage model [98], which assumes depassivated interface states contributing to the permanent part of NBTI (Fig. 9.8 (right)). In the simulation the permanent part was modeled by an additional defect level which can only be filled during stress. During recovery the defect level remains occupied.

A further issue when dealing with device simulations was the already mentioned exactness of the distributions due to the different amount of taken defects. Here a number of 1000 representative defects exhibits a good compromise between computational efforts and accuracy of the simulation when fitting the experiments. Therefore this value is chosen for the calibration of the parameter set. When taking more defects into account, the simulation results become smoother and do not contain the small kinks, as visible in Fig. 9.8 (right) for 1000 defects. However, this is only due to numerical reasons, since the actual degradation is always obtained by scaling the behavior of the “representative” defects. Consequently, the overall behavior is not changed, which can be seen when comparing Fig. 9.8 (right) and Fig. 9.9 (left).

At last, it can be pointed out that the classical approach, which assumes all holes to be energetically located at the valence band edge of the substrate, shows small deviations from the QM-results,

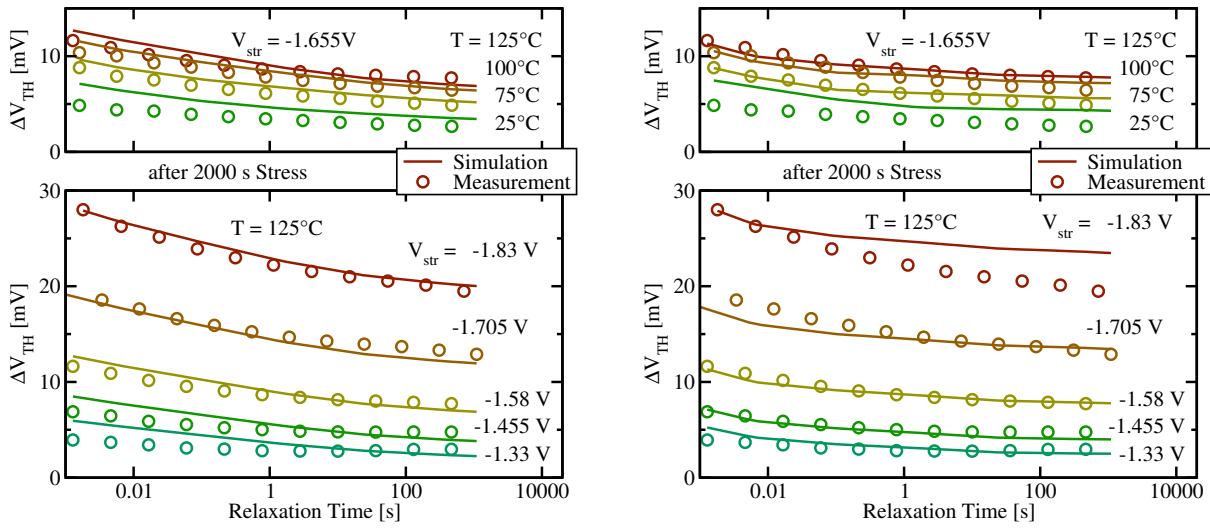


Figure 9.9: The last relaxation sequence as depicted in Fig. 9.8 (left) for different temperatures and stress voltages. **Left:** When using 10000 instead of 1000 defects the simulation results become smoother. Although the simulation was performed with the same parameter set, the measurement is quite well reproduced. **Right:** The classical results assume all holes to be energetically located at the valence band edge of the substrate. Except for the heaviest stress conditions (125°C and $V_{\text{str}} = -1.705\text{ V}$, -1.83 V) the classical approach can as well be applied and yields rather good agreement with the experiment.

especially for 125°C and $V_{\text{str}} = -1.83\text{ V}$, cf. Fig. 9.9 (right). This is due to the missing influence of the subbands which are localized in the SiGe-layer.

9.4 Conclusions

In this chapter it has been shown that the recovery of a multi-layer high-k pMOSFET structure after NBTI stress can be well modeled on the basis of a refined multi-state defect model, in which the effective rates of the model over two metastable states determine the temporal occupancy of a single defect. Since the area of the devices investigated here is large, also a large number of oxide defects having different defect quantities like energies, barriers inbetween, and positions inside the oxide are assumed. Due to the complexity of the device also quantum mechanical effects like subbands are incorporated, which is necessary to explain the heavier stress conditions. When the effective rates of all the defects for all the contributing subbands at certain times are summed up, the defect band occupation can be calculated giving the overall degradation. This is done for measurement data covering a large range of accelerated stress conditions yielding excellent agreement.

Chapter 10

Summary and Outlook

This work deals with the characterization of the bias temperature instability (BTI) from a metrological approach as well as from a modeling approach. To resolve the question “what to model”, well-designed measurements are of utmost importance. Numerous measurement techniques have therefore been compared and their applicability to BTI was carefully checked. Although it is seemingly obvious, issues like the measurement delay or the impact of the measurement setup on the device have been ignored for a long time. Just a few years ago the characterization of the negative BTI (NBTI) of pMOS devices as the most serious BTI condition was only focused on the stress phase, i.e. the degradation during BTI. This stress phase was widely modeled via the reaction diffusion theory which assumes the diffusion of hydrogen into the oxide. However, as soon as the stress is removed, a part of the degradation was found to recover, which can not be adequately described by the reaction diffusion theory and extended variants thereof.

The final failure of the long-reigning reaction-diffusion theory has demonstrated that without the use of significant and proper experimental NBTI data, modeling attempts are highly questionable, especially when they are based on incorrect premises, like the previously ignored recovery after NBTI. Upon the emergence of refined measurement techniques, many dependencies of the NBTI mechanism were explored, e.g. voltage, temperature, and frequency dependence to only mention a few, and new modeling attempts were published. While some attempts focused on the existence of interface states only, others assumed faster hole traps and slower interface states to contribute to NBTI. For the latter assumption it was suggested that the recovery behavior can be split into a recovering part due to hole traps and a permanent part due to interface states. In order to analyze the role of the recoverable component, measurement delay times down to the microsecond-regime became important.

The main experimental challenges discussed in this thesis include the synchronization of in- and output data with a high time resolution, the reduction of noise, and the clear distinction between stress phase and recovery phase when a pulsed measurement technique is used. Especially, the transition phase between stress and recovery is undefined. It is not surprising, however, that there is no perfect measurement technique and each of them has its specific drawbacks.

- On-the-fly (OTF) measurements allow a recovery-free measurement, since the drain current is measured at stress level. Its main drawback, however, lies in a missing unstressed reference and the fact that mobility degradation is wrongly attributed to the threshold voltage degradation.
- The charge pumping (CP) technique allows to assess the amount of interface states. Unfortunately, the characterization of the recovery behavior is extremely challenging because of the

inherent bias switches into accumulation. The extended OTF-CP technique, named OFIT, suffers from a spurious degradation during stress compared to that during recovery. Since this is due to the different pulse conditions, it can at least be corrected. However, the resolution of the short-term behavior is very limited due to the averaging of the gate pulses.

- A tool to access short-term stress as well as recovery was found in the fast pulsed $I_D(V_G)$ -characteristics. When carefully handled, delay times of a microsecond can be achieved. A general disadvantage of pulsed measurement techniques lies in the characterization of PBTI of pMOSFETs. Although stress is performed in the accumulation regime in this case, degradation has to be characterized in inversion anyway. As unwanted side effect this may yield small NBTI stress and hence obscure the real PBTI-induced degradation.
- Focusing on the recovery phase the measurement technique of choice in this thesis is the measurement-stress-measurement (MSM) technique. It allows the determination of an unstressed reference and exhibits very small delay times. Furthermore it can be extended to many alternating stress and recovery sequences, referred to as eMSM technique.

Using these measurement techniques the following results are obtained:

Irrespective of which stress condition (NBTI or PBTI) is imposed to which device type (nMOS or pMOS), always a negative shift of the threshold voltage is observed for thin oxides, with the largest degradation for the NBTI/pMOS case. Furthermore, the degradation can be split into a recoverable hole trap component and a permanent interface state component. When the recovery is monitored with a too large measurement delay time, a fraction of the recoverable part can be missed. Short-term NBTI stress and recovery measurements in the range of $1\ \mu\text{s}$ to $1\ \text{s}$ underline the challenge that the measurement delay time as well as the settling time of the applied gate pulse have a huge impact on the monitored recovery. A settled gate pulse for example may miss a part of the degradation and recovery which would yield spurious effects. For this reason a more liberal level for the pulse settling time should be used instead, involving the risk of mixing the stress and recovery phase.

To rigorously model BTI, it is important to not only study the short-term but also the long-term recovery behavior on an equal footing. New measurement results indicate that the previously assumed permanent part actually exhibits recovery, albeit on a large time scale. When the same type of pMOS previously subjected to NBTI stress is PBTI-stressed instead, a different recovery behavior is obtained. This can be explained by the assumption of a “general recovery behavior” which covers all aspects of the previous stress conditions. Depending on the oxide electric field and stress time only a part of the recovery is visible in the measurement. The invisible part of the general recovery is either too fast to be observable, which is assumed to be the case for NBTI, or so late that even an experiment lasting for two weeks is too short to capture the full recovery characteristics. The latter case is assumed for PBTI.

The fact that recovery even occurs below a microsecond and continues for at least weeks requires a model that is able to explain large time scales over more than 12 decades. Modelling attempts using Shockley-Read-Hall-like (SRH) processes are ruled out because they feature a too small time constant range. Also the field and temperature dependence can not be explained with the SRH model. The latest attempt is based on the non-radiative multi-phonon (NMP) theory, where hole capture and emission time constants depend on the barriers between two different defect states. Upon the application of stress, one defect configuration is shifted with respect to the other, which favors a hole capture process during stress. During recovery the energy levels between the two defect

levels favor hole emission. Depending on the kind of defect this gives a well-defined pair of a capture and emission time constant.

The step-like recovery behavior of small devices is a clear demonstration of such hole emission events from single defects. However, larger devices contain a larger number of defects. When in the simplest case a uniform distribution of time constants is assumed, a log-like recovery behavior consistent with the measurement results is obtained. This indicates that the underlying mechanism during BTI is actually based on the superposition of a large number of different defects that each feature different pairs of capture and emission times.

Finally, the recovery of high-k SiGe pMOSFETs was modeled by an extended NMP theory which is able to describe the behavior of switching oxide traps during NBTI: The multi-state defect model features two bistable defect states, each consisting of a stable and a metastable defect level. Effective capture and emission rates between the two stable defect levels determine the occupation of the defect. The complexity of the devices required the implementation of quantum mechanical effects like quantization in the channel to successfully model a large experimental dataset of different stress times, voltages and temperatures. To that end, a distribution of energies, barriers, and positions inside the oxide were assumed. The occupancies of all defects were finally summed up over all subbands. The overall degradation yielded excellent agreement with the experimental data, strongly indicating that the extended NMP theory is valid for NBTI.

However, the microscopic structure of the defect(s) contributing to NBTI and PBTI still remains vague. Recent publications have performed density function theory calculations using the oxygen vacancy and the hydrogen bridge as possible defect configurations. Unfortunately, the obtained thermodynamic energy levels of both defects are not in agreement with the experimental observations made during BTI. A successful defect identification would help the semiconductor industry to alter the manufacturing process in a manner that BTI would not be longer a serious reliability issue. Also, with the successful understanding of BTI it would be finally possible to make clearer predictions on the lifetime of MOSFETs.

Appendix A

Extracting V_θ Based on the Level 1 model

Different OTF models will be described below. Their advantages and disadvantages are due to different approximations [37]. The underlying compact model was already introduced in Section 2.3,

$$I_{D,\text{lin}} = \frac{\beta V_D (V_G - V_\theta - 1/2V_D)}{1 + \theta(V_G - V_\theta - 1/2V_D)}. \quad (\text{A.1})$$

While β depends on the effective mobility μ_{eff} , θ models the mobility saturation with increasing vertical field and V_θ the threshold voltage.

A.1 OTF1

The simplest OTF model, in the following named OTF1, assumes θ to be small, V_D to be small compared to V_G and V_θ , and μ_{eff} as constant ($\beta = 1$) [36, 49]. Equation (A.1) then simplifies to

$$I_{D,\text{lin}} = \beta V_D (V_G - V_\theta). \quad (\text{A.2})$$

With only V_θ contributing to the total differential

$$dI_{D,\text{lin}} = \frac{\partial I_{D,\text{lin}}}{\partial V_\theta} dV_\theta + \frac{\partial I_{D,\text{lin}}}{\partial \beta} d\beta + \frac{\partial I_{D,\text{lin}}}{\partial V_G} dV_G + \dots$$

and by approximating $\Delta I_{D,\text{lin}} \approx dI_{D,\text{lin}}$ yields

$$\Delta I_{D,\text{lin}} = \underbrace{\frac{\partial I_{D,\text{lin}}}{\partial V_\theta}}_{-\beta V_D} \Delta V_\theta = - \underbrace{\frac{\partial I_{D,\text{lin}}}{\partial V_G}}_{g_m} \Delta V_\theta. \quad (\text{A.3})$$

Equating (A.2) with (A.3) finally gives

$$\Delta V_\theta^{\text{OTF},1} \approx - \frac{\Delta I_{D,\text{lin}}}{I_{D,\text{lin}}} (V_G - V_\theta) = - \frac{\Delta I_{D,\text{lin}}}{g_m}. \quad (\text{A.4})$$

Note the similarity to (2.9) in (A.4). Hence, OTF1 only requires the determination $I_{D,\text{lin}}$ during stress but neglects the mobility reduction described by β and θ .

A.2 OTF2

Starting again with (A.1), but this time without simplifications, the total differential delivers

$$g_m = \frac{\partial I_{D,\text{lin}}}{\partial V_G} = -\frac{\partial I_{D,\text{lin}}}{\partial V_\theta} = -\frac{-\beta V_D}{(1 + \theta(V_G - V_\theta - 1/2V_D))^2}. \quad (\text{A.5})$$

Linking (A.1) and (A.5) results in

$$\frac{I_{D,\text{lin}}}{g_m} = (1 + \theta(V_G - V_\theta - 1/2V_D))(V_G - V_\theta - 1/2V_D) \quad (\text{A.6})$$

which is needed for the mobility variation. Differentiating this expression with respect to V_θ describes ΔV_θ as a function of the measured change in $I_{D,\text{lin}}/g_m$ not depending on β [6].

$$\frac{\partial(\frac{I_{D,\text{lin}}}{g_m})}{\partial V_\theta} = -1 - 2\theta(V_G - V_\theta - 1/2V_D) \quad (\text{A.7})$$

$$\Delta V_\theta^{\text{OTF},2} \approx \frac{-\Delta(\frac{I_{D,\text{lin}}}{g_m})}{1 + 2\theta(V_G - V_\theta - 1/2V_D)} \quad (\text{A.8})$$

In contrast to OTF1, OTF2 requires a complete $I_D(V_G)$ -characteristics for the extraction of θ .

A.3 OTF3

The OTF3-method is explained in Chapter 2.3.

Appendix B

Ideal MOS Capacitor

The band diagrams of an ideal MOS structure consisting of a gate electrode (metal or polysilicon), a dielectric (oxide), and a semiconductor (nMOS or pMOS) are shown in Fig. B.1 under different operating conditions for both nMOS and pMOS. For the most simple case it is assumed that (i) there are no charges in the oxide, (ii) the resistivity of the oxide is infinite, and (iii) the work function difference between the metal and the semiconductor, ϕ_{ms} , is zero [10]. The operating conditions depend on the applied voltage V on the metal contact with respect to the Fermi level of the grounded semiconductor and are called accumulation (a), flatband (b), depletion (c), and inversion (d).

In the following the pMOSFET with n-substrate will be explained:

(a) When a positive voltage is applied at the contact the conduction band E_c bends down towards the Fermi level E_f that is set constant in the semiconductor where no current flows. This bending yields an accumulation of the majority carriers (electrons) near the interface.

(b) For $V = 0$ all bands remain flat and the semiconductor and its majority and minority carriers are in thermal equilibrium.

(c) Under a small negative voltage the majority carriers are repelled from the interface, involving that the bands are bend up. The intrinsic energy E_i gets closer to E_f .

(d) When further increasing the negative voltage this bending continues and once E_i crosses E_f the minority carriers (holes) exceed the majority carriers at the interface. Hence, this case is called inversion, as the interface is inverted.

For the p-type structure with holes as majority carriers and electrons as minority carriers only the polarity of the voltage has to be changed.

B.1 Surface Space Charge Region of an n-Type MOS Capacitor

To be able to calculate and interpret the $C(V)$ -characteristics of pMOSFETs the understanding of the semiconductor charge Q_s as a function of the surface potential ψ_s is inevitable. The nomenclature of these calculations is based on [10] but an n-type instead of a p-type semiconductor is used.

The measure for the local band bending in the semiconductor is given by $\psi_n(x) = E_i(x)/q$, which determines the local electron concentration $n_n(x)$ and hole concentration $p_n(x)$ given by

$$n_n(x) = n_{n0} \exp\left(\frac{q\psi_n}{k_B T}\right) = n_{n0} \exp(\beta\psi_n) \quad (\text{B.1})$$

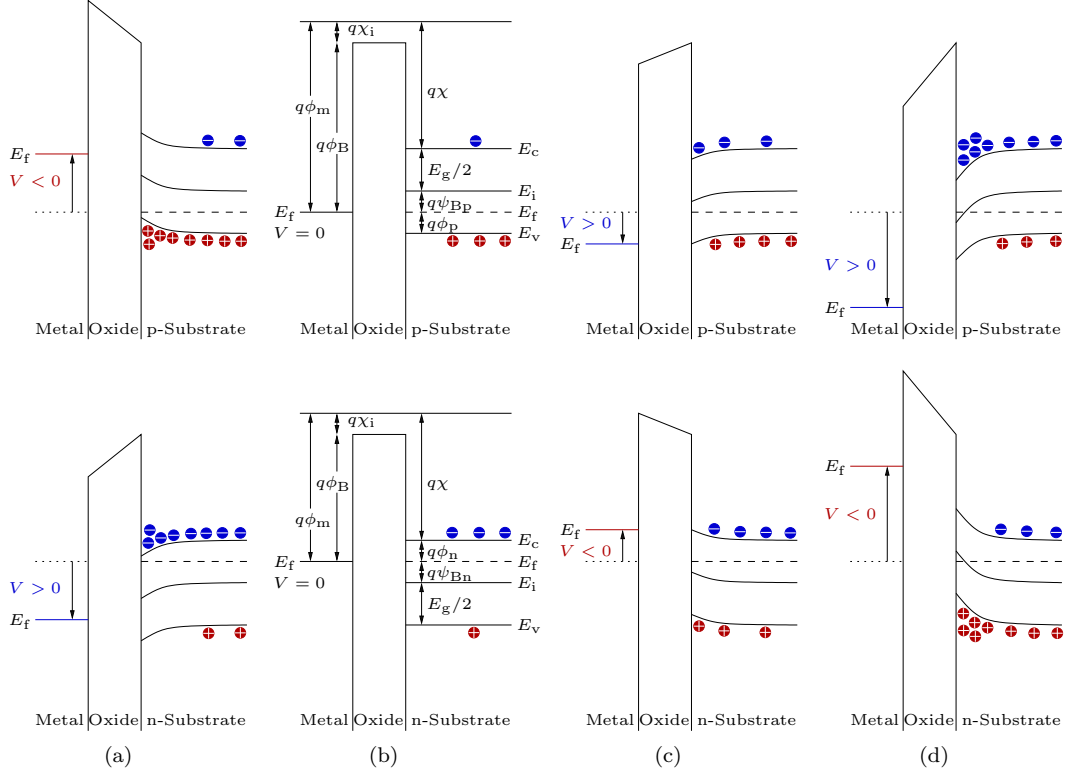


Figure B.1: The energy band diagrams for ideal MOS-capacitors under different bias conditions: (a) accumulation, (b) flatband, (c) depletion, and (d) inversion. The resulting charges bend the bands upwards near the interface of the oxide/substrate (insulator/semiconductor) if $V < 0$ and downwards if $V > 0$. The energy levels and potentials are marked for the flatband condition ($V = 0$), with ϕ_{ms} denoting the Fermi potential with respect to the vacuum level, χ_i and χ , as electron affinity for the oxide and the substrate, and E_g as bandgap in the substrate. **Top:** For a p-semiconductor (nMOS) it holds that $\phi_{ms} \equiv \phi_m - (\chi + E_g/q - \phi_p) = 0$, where ψ_{BP} and ϕ_p represent the Fermi potentials with respect to the intrinsic energy E_i and valence band E_v . **Bottom:** For an n-semiconductor (pMOS) one obtains $\phi_{ms} \equiv \phi_m - (\chi + \phi_n) = 0$ with ϕ_n and ψ_{Bn} as the Fermi potentials with respect to the conduction band E_c and intrinsic energy E_i .

$$p_n(x) = p_{n0} \exp\left(\frac{-q\psi_n}{k_B T}\right) = p_{n0} \exp(-\beta\psi_n) \quad (\text{B.2})$$

where n_{n0} and p_{n0} denote the equilibrium densities (flatband case), k_B the Boltzmann constant and $1/\beta$ the thermoelectrical potential at T . Starting from the 1d-Poisson equation

$$\frac{d^2\psi_n(x)}{dx^2} = -\frac{\rho(x)}{\epsilon_1\epsilon_0} \quad (\text{B.3})$$

with the local space charge density $\rho(x) = q(N_D^+ - N_A^- + p_n - n_n)$ the potential is obtained. Here, N_D^+ and N_A^- are the densities of the ionized donors and acceptors. Deep in the substrate (and under

flatband conditions in the whole semiconductor) charge neutrality can be assumed $\rho(x) = 0$ due to $\psi_n(x \rightarrow \infty) = 0$. Inserting and evaluating (B.1) and (B.2) yields $N_D^+ - N_A^- = n_{n0} - p_{n0}$ and finally

$$\begin{aligned} \frac{d^2\psi_n(x)}{dx^2} &= -\frac{q}{\epsilon_r\epsilon_0}(n_{n0} - p_{n0} + p_n - n_n) \\ &= -\frac{q}{\epsilon_r\epsilon_0}(p_{n0}(\exp(-\beta\psi_n) - 1) - n_{n0}(\exp(\beta\psi_n) - 1)). \end{aligned} \quad (\text{B.4})$$

To integrate (B.4) the following integration trick is necessary:

$$\begin{aligned} d\left(\frac{d\psi}{dx}\right) &= \frac{d^2\psi dx - d\psi d^2x}{dx^2} = \frac{d^2\psi}{dx} \\ \frac{d\psi}{dx} d\left(\frac{d\psi}{dx}\right) &= \frac{d\psi}{dx} \frac{d^2\psi}{dx} \\ &= d\psi \frac{d^2\psi}{dx^2} \\ \int \frac{d\psi}{dx} d\left(\frac{d\psi}{dx}\right) &= \int \frac{d^2\psi}{dx^2} d\psi \end{aligned} \quad (\text{B.5})$$

With (B.5) the Poisson equation (B.4) can be rewritten as

$$\begin{aligned} \int_0^{-E_s} \frac{d\psi_n}{dx} d\left(\frac{d\psi_n}{dx}\right) &= \int_0^{\psi_s} -\frac{q}{\epsilon_r\epsilon_0}(p_{n0}(\exp(-\beta\psi_n) - 1) - n_{n0}(\exp(\beta\psi_n) - 1)) d\psi_n \\ \frac{1}{2} \left(\frac{d\psi_n}{dx}\right)^2 \Big|_0^{-E_s} &= -\frac{q}{\epsilon_r\epsilon_0} \left[p_{n0} \left(\frac{\exp(-\beta\psi_n)}{-\beta} - \psi_n \right) - n_{n0} \left(\frac{\exp(\beta\psi_n)}{\beta} - \psi_n \right) \right] \Big|_0^{\psi_s} \end{aligned} \quad (\text{B.6})$$

$$\begin{aligned} \frac{1}{2} E_s^2 &= -\frac{q}{\epsilon_r\epsilon_0} \left[p_{n0} \left(\frac{\exp(-\beta\psi_s)}{-\beta} - \psi_s + \frac{1}{\beta} \right) - n_{n0} \left(\frac{\exp(\beta\psi_s)}{\beta} - \psi_s - \frac{1}{\beta} \right) \right] \\ &= \frac{qn_{n0}}{\epsilon_r\epsilon_0\beta} \left[\left(\frac{p_{n0}}{n_{n0}} \right) (\exp(-\beta\psi_s) + \beta\psi_s - 1) + (\exp(\beta\psi_s) - \beta\psi_s - 1) \right]. \end{aligned} \quad (\text{B.7})$$

The integration boundaries (B.6) range from the substrate, where $\psi_n = 0$ and $d\psi_n/dx = 0$ to the surface with $\psi_n = \psi_s$ and $d\psi_n/dx = -E_s$.

For non-degenerate semiconductors the Fermi level is far enough away from E_c and E_v and Fermi-Dirac statistics can be approximated by Boltzmann statistics. The carrier concentrations for an n-type semiconductor with $N_D^+ > N_A^-$ can be approximated [10]. Then,

$$n_{n0} \approx N_D = n_i \exp\left(\frac{E_f - E_i}{k_B T}\right) = n_i \exp\left(\frac{q\psi_{Bn}}{k_B T}\right) \quad (\text{B.8})$$

$$p_{n0} = \frac{n_i^2}{n_{n0}} \approx \frac{n_i^2}{N_D}, \quad (\text{B.9})$$

which yields

$$\frac{p_{n0}}{n_{n0}} = \exp(-2\beta\psi_{Bn}). \quad (\text{B.10})$$

The electric field at the surface in (B.7) is now simplified to

$$E_s = \pm \sqrt{\frac{2qn_{n0}}{\epsilon_r\epsilon_0\beta}} F(\beta\psi_{Bn}, \psi_s) \quad (\text{B.11})$$

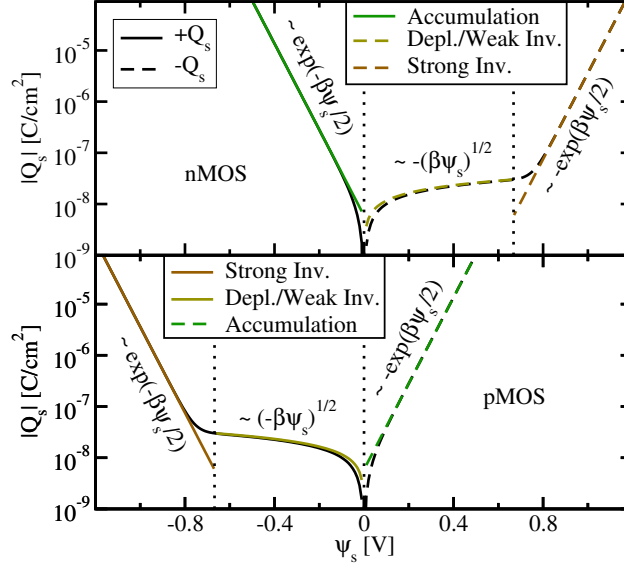


Figure B.2: The surface charge density Q_s compared for both p-type and n-type semiconductors depending on ψ_s . While the solid lines stand for positive Q_s , the dashed lines symbolize a negative Q_s . The approximations very well fit the exact solutions (B.13) and (B.18) drawn in black. Deviations from the latter only exist at the transitions of the operating regimes. It is shown that $Q_{s,n}(\psi_s) \equiv -Q_{s,p}(-\psi_s)$, where the subscripts n and p denote the type of semiconductor.

with

$$F(\beta\psi_{Bn}, \psi_s) = +\sqrt{\exp(-2\beta\psi_{Bn})(\exp(-\beta\psi_s) + \beta\psi_s - 1) + (\exp(\beta\psi_s) - \beta\psi_s - 1)}. \quad (\text{B.12})$$

By applying Gauss' law $\nabla \cdot E = \rho(x)/(\epsilon_r \epsilon_0)$ the space-charge-density per area is finally obtained as

$$Q_s = -\epsilon_r \epsilon_0 E_s = \mp \sqrt{\frac{2q\epsilon_r \epsilon_0 n_{n0}}{\beta}} F(\beta\psi_{Bn}, \psi_s). \quad (\text{B.13})$$

For n-type (pMOS) semiconductors (B.13) can now be approximated for the different regimes of the surface potential $\psi_s \equiv V$ with the contact voltage V , defined in the beginning of the chapter. For accumulation with $\psi_s > 0$, the term $\exp(\beta\psi_s)$ dominates in (B.12), making

$$Q_s \propto -\exp(\beta\psi_s/2).$$

For $\psi_s < 0$, depletion and successively weak inversion set in till $\psi_s = -2\psi_{Bn}$ is fulfilled. Here

$$Q_s \propto +\sqrt{-\beta\psi_s}.$$

Finally, beyond $\psi_s > -|2\psi_{Bn}|$ the first term in (B.12) starts to dominate, which yields

$$Q_s \propto +\exp(-\beta\psi_s/2).$$

B.2 Results for p-Type Semiconductors

When compared to n-type semiconductors, some of the equations from (B.1) to (B.13) differ for p-type semiconductors. Due to their exchanged concentrations of holes p_p and electrons n_p

$$p_p(x) = p_{p0} \exp(-\beta\psi_p) \quad p_{p0} \approx N_A = n_i \exp(\beta\psi_{Bp}) \quad (\text{B.14})$$

$$n_p(x) = n_{p0} \exp(\beta\psi_p) \quad n_{p0} = n_i^2/p_{p0} \approx n_i^2/N_A \quad (\text{B.15})$$

some of the equations change their signs depending on the applied bias conditions. The calculations are derived in [10] and for that reason only the differences between n-type and p-type are summarized below. Starting with the electric field at the interface

$$\frac{1}{2}E_s^2 = \frac{qp_{p0}}{\epsilon_r\epsilon_0\beta} \left[(\exp(-\beta\psi_s) + \beta\psi_s - 1) + \left(\frac{n_{p0}}{p_{p0}} \right) (\exp(\beta\psi_s) - \beta\psi_s - 1) \right]$$

$$E_s = \pm \sqrt{\frac{2qp_{p0}}{\epsilon_r\epsilon_0\beta}} F(\beta\psi_{Bp}, \psi_s) \quad (\text{B.16})$$

with

$$F(\beta\psi_{Bp}, \psi_s) = +\sqrt{(\exp(-\beta\psi_s) + \beta\psi_s - 1) + \exp(-2\beta\psi_{Bp})(\exp(\beta\psi_s) - \beta\psi_s - 1)}, \quad (\text{B.17})$$

the space-charge-density becomes

$$Q_s = -\epsilon_r\epsilon_0 E_s = \mp \sqrt{\frac{2q\epsilon_r\epsilon_0 p_{p0}}{\beta}} F(\beta\psi_{Bp}, \psi_s). \quad (\text{B.18})$$

Again, the charge at the surface (B.18) can be approximated for certain surface potentials ψ_s .

For accumulation with $\psi_s < 0$, the term $\exp(-\beta\psi_s)$ dominates the root in (B.17), making $Q_s \propto \exp(-\beta\psi_s/2)$.

Starting from the flatband condition at $\psi_s = 0$, first depletion of holes and afterwards weak inversion set in till $\psi_s = 2\psi_{Bp}$ is fulfilled. In these two regimes $Q_s \propto -\sqrt{\beta\psi_s}$.

Finally, beyond $\psi_s > 2\psi_{Bn}$ the first term in (B.17), starts to dominate by outbalancing the negative exponent in $\exp(-2\beta\psi_{Bp})$ which yields $Q_s \propto -\exp(\beta\psi_s/2)$.

In Fig. B.2 the different operating conditions with its resulting surface charge density Q_s at the interface side of the semiconductor are opposite for both p-type and n-type semiconductors. The above mentioned approximations very well fit the exact solutions (B.13) and (B.18), as deviations are only present at the intersections of the different regimes. Furthermore, it is shown that $Q_{s,n}(\psi_s) \equiv -Q_{s,p}(-\psi_s)$, where the subscripts n and p denote the type of semiconductor.

Appendix C

Diffusion-Limited Stress Phase of the Reaction-Diffusion Theory

In the standard reaction-diffusion theory the kinetic rate equation describing the interface reaction via a hydrogen species X_{it} [56–58] in

$$\frac{\partial N_{it}}{\partial t} = k_f(N_0 - N_{it}) - k_r N_{it} X_{it}^{1/a}, \quad (C.1)$$

with N_{it} and N_0 as unpassivated and total amount of interface states. Thus $N_0 - N_{it}$ denotes the concentration of passivated interface defects de-passivating with the rate k_f . The passivation rate k_r of the dangling bonds also depends on the hydrogen species X_{it} with its kinetic exponent a (1 for H^0 and H^+ , and 2 for H_2) [172]. Assuming the quasi-equilibrium regime of the interface reaction ($\partial N_{it}/\partial t \approx 0$) as the dominant regime after [17, 59, 66, 71], the rate equation (C.1) can be rewritten as

$$X_{it} = \left(\frac{k_f N_0 - N_{it}}{k_r N_{it}} \right)^a. \quad (C.2)$$

The boundary value problem for X_{it} is as follows:

$$\frac{\partial X_{it}(x, t)}{\partial t} = D_X \frac{\partial^2 X_{it}(x, t)}{\partial x^2} \pm \mu E_{ox} \frac{\partial X_{it}(x, t)}{\partial x}. \quad (C.3)$$

When neglecting charged hydrogen (H^+), the drift term inside the drift-diffusion process (C.3) vanishes. The remaining diffusion process will be approximately solved on basis of a triangular hydrogen profile¹ [59, 85], as depicted in Fig. C.1 (right).

After the continuity equation the interface states additionally created, $\Delta N_{it}(t)$, are due to the leaving hydrogen $X_{it}(t)$, which is shown in Fig. C.1 (middle). The corresponding diffusion front is given by $\sqrt{D_X t}$. Comparing with Fig. C.1 (right) yields the area confined by the diffusion front on the one hand and the number of aX_{it} at the interface at the other hand which equals

$$\frac{aX_{it}(t)\sqrt{D_X t}}{2} = \Delta N_{it}(t). \quad (C.4)$$

The ratio between the diffusing hydrogen species and resulting interface states is determined by the kinetic exponent a , i.e. each H_2 leaves 2 dangling bonds.

¹Other diffusion profiles, e.g. a Gaussian profile, are more accurate, but do not change the overall result [173].

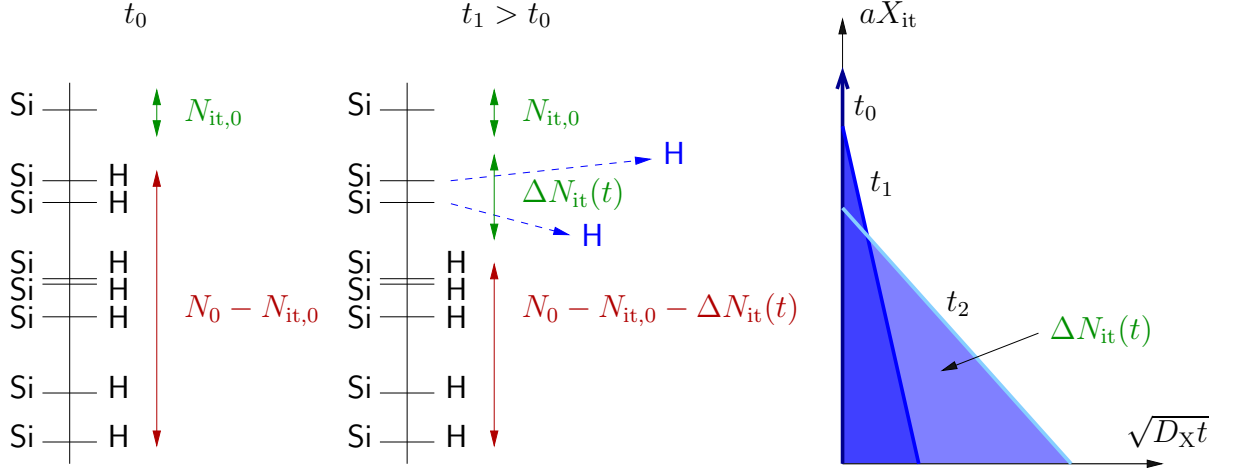


Figure C.1: **Left:** For the diffusion-limited part of the RD theory the interface is considered to be nearly fully passivated at t_0 , i.e. $N_0 \gg N_{it,0}$. The additionally created dangling bonds ΔN_{it} are furthermore assumed to dominate the total number of dangling bonds ($N_{it,0} + \Delta N_{it}(t) \approx \Delta N_{it}(t)$). Note that in fact the Si-H and the silicon dangling bonds are not arranged in two groups like schematically depicted here, but are randomly distributed. **Right:** Hydrogen profile inside the oxide during a diffusion process. The area under the hydrogen profile with its progressing diffusion front $\sqrt{D_X t}$ equals the number of additionally generated interface states $\Delta N_{it}(t)$ given by relation (C.4).

To solve equations (C.2) and (C.4) the following assumptions are made: (i) The amount of passivated interface states N_0 is much larger than the initial value of N_{it} , $N_{it,0}$, and (ii) $\Delta N_{it}(t) \gg N_{it,0}$. A schematic picture of the interface shows all necessary quantities and its relations (Fig. C.1 (left)). Inserting these assumptions into (C.2) and comparing with (C.4) gives

$$\frac{2\Delta N_{it}(t)}{a\sqrt{D_X t}} \approx \left(\frac{k_f}{k_r} \frac{N_0}{\Delta N_{it}(t)} \right)^a. \quad (C.5)$$

The approximated number of ΔN_{it} is then

$$\Delta N_{it}(t) = \left(\frac{k_f}{k_r} N_0 \right)^{\frac{a}{a+1}} \left(\frac{a}{2} \right)^{\frac{1}{a+1}} (D_X t)^{\frac{1}{2(a+1)}}. \quad (C.6)$$

By using atomic hydrogen ($a=1$) this term simplifies to

$$\Delta N_{it}(t) = \sqrt{\frac{k_f N_0}{k_r}} (D_X t)^{\frac{1}{4}}, \quad (C.7)$$

while molecular hydrogen ($a=2$) yields

$$\Delta N_{it}(t) = \left(\frac{k_f}{k_r} N_0 \right)^{\frac{2}{3}} (D_X t)^{\frac{1}{6}}. \quad (C.8)$$

Alternatively (C.4) can be formulated via the flux of the hydrogen profile (the gradient right at the interface) and yields a first-order differential equation in time to solve. The results differing by a constant prefactor from the algebraic expressions in (C.6) are summarized in [71].

Appendix D

Multi-Phonon Emission

Based on Chapter 8.3 the forward and reverse barriers of the radiative multi-phonon emission (MPE) and the non-radiative multi-phonon (NMP) model will be deduced from the harmonic oscillator of the form

$$V_1(q) = \frac{1}{2}M\omega_1^2(q - q_1)^2 + E_1 \quad (\text{D.1})$$

$$V_2(q) = \frac{1}{2}M\omega_2^2(q - q_2)^2 + E_2, \quad (\text{D.2})$$

where E_1 and E_2 denote the defect states in thermal equilibrium at $V_1(q_1)$ and $V_2(q_2)$. The vibronic frequencies ω_1 and ω_2 set the curvature of the harmonic potentials.

D.1 Radiative Multi-Phonon Emission

In a radiative process the energy necessary for a transition from V_1 to V_2 is obtained from the binding energy, $E_B(q) = V_2(q) - V_1(q)$. This energy writes as

$$\begin{aligned} E_B(q) &= \underbrace{E_2 - E_1}_{E_{21}} + \frac{1}{2}M\omega_2^2(q - q_2)^2 - \frac{1}{2}M\omega_1^2(q - q_1)^2 \\ &= E_{21} + \frac{1}{2}M\omega_2^2 \left(\underbrace{(q_2^2 - 2q_1q_2 + q_1^2)}_{(q_1 - q_2)^2} + 2q_1q_2 - q_1^2 - 2qq_2 + q^2 \right) - \underbrace{\frac{\omega_1^2}{\omega_2^2}}_{R^2} (q - q_1)^2 \\ &= E_{21} + S\hbar\omega_2 + \frac{1}{2}M\omega_2^2 \left((2q_1q_2 - q_1^2 - 2qq_2 + q^2) - R^2 (q - q_1)^2 \right), \end{aligned} \quad (\text{D.3})$$

with the relaxation energy $S\hbar\omega_2 = \frac{1}{2}M\omega_2^2(q_1 - q_2)^2$ from $V_2(q_1)$ to $V_2(q_2)$. The two points where a radiative transition is possible after the Franck-Condon principle are at q_1 and q_2 . Inserting yields

$$\begin{aligned} E_B(q_1) &= \epsilon_{12} = E_{21} + S\hbar\omega_2 \\ E_B(q_2) &= \epsilon_{21} = E_{21} + S\hbar\omega_2 + \underbrace{\frac{1}{2}M\omega_2^2 \left(\underbrace{-q_1^2 + 2q_1q_2 - q_2^2}_{-(q_1 - q_2)^2} \right)}_0 - \frac{1}{2}M\omega_1^2 (q - q_1)^2 \\ \epsilon_{21} &= E_{21} - S\hbar\omega_1. \end{aligned} \quad (\text{D.4})$$

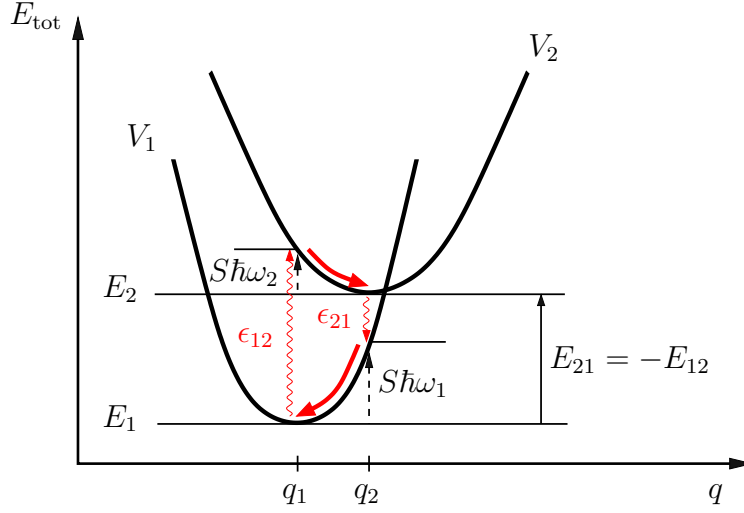


Figure D.1: Description of a radiative multi-phonon emission process assuming harmonic oscillators with different vibronic frequencies ω_1 and ω_2 in a reaction coordinate diagram. The photon energy required to change from V_1 to V_2 equals $\epsilon_{12} = E_{21} + S\hbar\omega_2$. Due to structural relaxation, the photon emitted in the following reverse process is smaller, namely $\epsilon_{21} = E_{21} - S\hbar\omega_1$.

Analogously to $S\hbar\omega_2$, $S\hbar\omega_1$ is defined as the relaxation energy from $V_1(q_2)$ to $V_1(q_1)$. A full MPE process is schematically depicted in Fig. D.1 for quadratic coupling ($\omega_1 \neq \omega_2$). In the case of linear coupling ($\omega_1 = \omega_2$) both relaxation energies coincide.

D.2 Non-Radiative Multi-Phonon Process

When there are no photons available for the transition, the process is called non-radiative multi-phonon (NMP) process. Now the transition energy from one parabolic minimum into the other has to be provided by phonons. Due to energy conservation, a classical transition at the points is possible, where the binding energy is zero. This is the case only at the intersection point¹ IP of (D.1) and (D.2). The value between $V_1(q_{\text{IP}}) = V_2(q_{\text{IP}})$ and either E_1 or E_2 then specifies the classical barrier which has to be crossed. Assuming linear coupling, i.e. $\omega_1 = \omega_2 = \omega$, yields

$$q_{\text{IP}} = \frac{\frac{2(E_1 - E_2)}{M\omega^2} + q_1^2 - q_2^2}{2(q_1 - q_2)}. \quad (\text{D.5})$$

Reinserting (D.5) into (D.2) delivers

$$\begin{aligned} V_2(q_{\text{IP}}) &= \frac{1}{2}M\omega^2 \left(\frac{\frac{2(E_1 - E_2)}{M\omega^2} + q_1^2 - q_2^2}{2(q_1 - q_2)} - q_2 \right)^2 + E_2 \\ &= \frac{1}{2}M\omega^2 \left(\frac{2(E_1 - E_2) + M\omega^2(q_1^2 - 2q_1q_2 + q_2^2)}{2M\omega^2(q_1 - q_2)} \right)^2 + E_2 \\ &= \frac{1}{2}M\omega^2 \left(\frac{2E_{12} + M\omega^2(q_1 - q_2)^2}{2M\omega^2(q_1 - q_2)} \right)^2 + E_2. \end{aligned} \quad (\text{D.6})$$

¹Assuming equal vibronic frequencies $\omega_1 = \omega_2$ yields only one, while $\omega_1 \neq \omega_2$ either yields two or no IP at all.

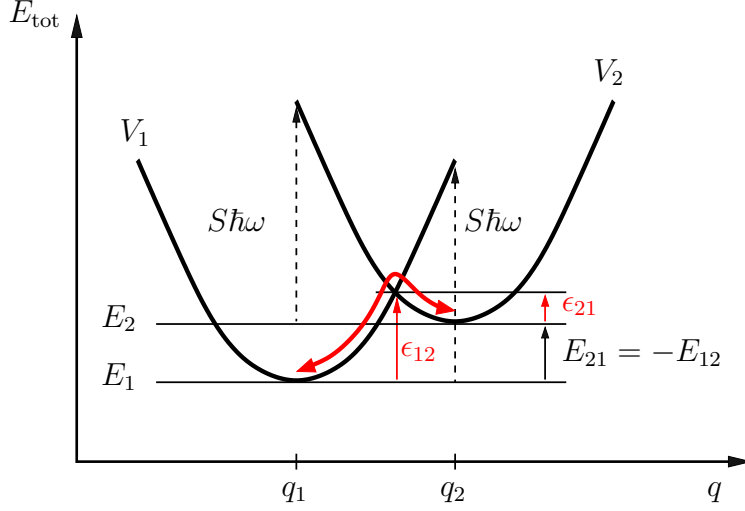


Figure D.2: Description of defect state transitions by assuming harmonic oscillators with equal vibronic frequency ω in a reaction coordinate diagram for the NMP model. The transition from one defect system into another happens at the intersection of the energy curves. Usually one transition is favored, inducing a preferred defect state. In this case the state $V_1(q_1)$ will be occupied most of the time.

With the relaxation energy $S\hbar\omega = \frac{1}{2}M\omega^2(q_1 - q_2)^2$ (cf. Fig. D.2), (D.6) can be further evaluated to

$$V_2(q_{\text{IP}}) = \frac{1}{2}M\omega^2 \frac{(2E_{12} + 2S\hbar\omega)^2}{4M^2\omega^4(q_1 - q_2)^2} + E_2 = \frac{(E_{12} + S\hbar\omega)^2}{4S\hbar\omega} + E_2. \quad (\text{D.7})$$

By combining (D.5) and (D.1), $V_1(q_{\text{IP}})$ can be written as

$$V_1(q_{\text{IP}}) = \frac{(E_{12} - S\hbar\omega)^2}{4S\hbar\omega} + E_1. \quad (\text{D.8})$$

The forward and reverse rates then read

$$\epsilon_{12} = \frac{(E_{12} - S\hbar\omega)^2}{4S\hbar\omega} = \frac{(S\hbar\omega + E_{21})^2}{4S\hbar\omega}, \quad (\text{D.9})$$

$$\epsilon_{21} = \frac{(E_{12} + S\hbar\omega)^2}{4S\hbar\omega} = \frac{(S\hbar\omega - E_{21})^2}{4S\hbar\omega}. \quad (\text{D.10})$$

In Fig. D.2 all derived quantities are depicted. Again S gives the number of emitted phonons with an energy of $\hbar\omega$. Their product $S\hbar\omega$ reflects the strength of coupling and hence has a huge impact on the transition rates of the defect states, e.g. a smaller $S\hbar\omega$ yields a smaller barrier.

Bibliography

- [1] J.E. Lilienfeld, “Method and apparatus for controlling electric current,” US-patent 1745175.
- [2] R.H. Dennard, F.H. Gaensslen, H.-N. Yu, V.L. Rideout, E. Bassous, and A.R. Leblanc, “Design of Ion-Implanted MOSFET’s with Very Small Physical Dimensions,” *IEEE J.Solid-State Circuits*, vol. 9, no. 5, pp. 256–268, October 1974.
- [3] Y. Miura and Y. Matukura, “Investigation of Silicon-Silicon Dioxide Interface Using MOS Structure,” *Japanese Journal of Applied Physics*, vol. 5, no. 2, pp. 180, 1966.
- [4] B.E. Deal, M. Sklar, A.S. Grove, and E.H. Snow, “Characteristics of the Surface-State Charge ($Q_{\{ss\}}$) of Thermally Oxidized Silicon,” *J.Electrochem.Soc.*, vol. 114, no. 3, pp. 266–274, 1967.
- [5] A.H. Edwards, “Theory of the P_b Center at the $\langle 111 \rangle$ Si/SiO₂ Interface,” *Physical Review B*, vol. 36, no. 18, pp. 9638–9648, December 1987.
- [6] V. Huard, M. Denais, and C. Parthasarathy, “NBTI Degradation: From Physical Mechanisms to Modelling,” *Microelectronics Reliability*, vol. 46, no. 1, pp. 1–23, 2006.
- [7] D.K. Schroder and J.A. Babcock, “Negative Bias Temperature Instability: Road to Cross in Deep Submicron Silicon Semiconductor Manufacturing,” *Journal of Applied Physics*, vol. 94, no. 1, pp. 1–18, Jul. 2003.
- [8] J.H. Stathis and S. Zafar, “The Negative Bias Temperature Instability in MOS Devices: A Review,” *Microelectronics Reliability*, vol. 46, no. 2-4, pp. 270, 2006.
- [9] V. Reddy, A. T. Krishnan, A. Marshall, J. Rodriguez, S. Natarajan, T. Rost, and S. Krishnan, “Impact of Negative Bias Temperature Instability on Digital Circuit Reliability,” in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2002, pp. 248–254.
- [10] S.M. Sze and K.K. Ng, *Physics of Semiconductor Devices*, John Wiley & Sons - Interscience, 3rd edition, 2007.
- [11] H. Reisinger, O. Blank, W. Heinrigs, A. Mühlhoff, W. Gustin, and C. Schlünder, “Analysis of NBTI Degradation- and Recovery-Behavior Based on Ultra Fast V_{th} -Measurements,” in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2006, pp. 448–453.
- [12] H. Reisinger, U. Brunner, W. Heinrigs, W. Gustin, and C. Schlünder, “A Comparison of Fast Methods for Measuring NBTI Degradation,” *IEEE Transactions on Device and Materials Reliability*, vol. 7, no. 4, pp. 531–539, Dec. 2007.

- [13] K. Jeppson and C. Svensson, "Negative Bias Stress of MOS Devices at High Electric Fields and Degradation of MNOS Devices," *Journal of Applied Physics*, vol. 48, no. 5, pp. 2004–2014, 1977.
- [14] S. Rangan, N. Mielke, and E.C.C. Yeh, "Universal Recovery Behavior of Negative Bias Temperature Instability," in *Proceedings of IEEE International Electron Devices Meeting (IEDM)*, 2003, pp. 341–344.
- [15] H. Reisinger, O. Blank, W. Heinrigs, W. Gustin, and C. Schlünder, "A Comparison of Very Fast to Very Slow Components in Degradation and Recovery due to NBTI and Bulk Hole Trapping to Existing Physical Models," *IEEE Transactions on Device and Materials Reliability*, vol. 7, no. 1, pp. 119–129, Mar. 2007.
- [16] A. Ortiz-Conde, F.J. García Sánchez, J.J. Liou, A. Cerdeira, M. Estrada, and Y. Yue, "A Review of Recent MOSFET Threshold Voltage Extraction Methods," *Microelectronics Reliability*, vol. 42, pp. 583–596, 2002.
- [17] B. Kaczer, V. Arkhipov, R. Degraeve, N. Collaert, G. Groeseneken, and M. Goodwin, "Disorder-Controlled-Kinetics Model for Negative Bias Temperature Instability and its Experimental Verification," in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2005, pp. 381–387.
- [18] B. Kaczer, T. Grasser, Ph.J. Roussel, J. Martin Martinez, R. O'Connor, B.J. O'Sullivan, and G. Groeseneken, "Ubiquitous Relaxation in BTI Stressing New Evaluation and Insights," in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2008, pp. 20–27.
- [19] A. Kerber, E. Cartier, L. Pantisano, M. Rosmeulen, R. Degraeve, T. Kauerauf, G. Groeseneken, H.E. Maes, and U. Schalke, "Characterization of the VT-Instability in SiO₂/HfO₂ Gate Dielectrics," in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2003, pp. 41–45.
- [20] C. Shen, M.-F. Li, X.P. Wang, Y.-C. Yeo, and D.-L. Kwong, "A Fast Measurement Technique of MOSFET $I_D(V_G)$ Characteristics," *IEEE Electron Device Letters*, vol. 27, no. 1, pp. 55–57, 2006.
- [21] D. Heh, R. Choi, C. D. Young, and G. Bersuker, "Fast and Slow Charge Trapping/Detrapping Processes in High-k nMOSFETs," in *Proceedings of IEEE International Integrated Reliability Workshop*, 2006, pp. 120–124.
- [22] J. P. Campbell, K. P. Cheung, J. S. Suehle, and A. S. Oates, "New Insight into NBTI Transient Behavior Observed from Fast- G_M Measurements," *IEEE Electron Device Letters*, vol. 29, no. 9, pp. 1065–1067, September 2008.
- [23] D. Heh, R. Choi, C. D. Young, B. H. Lee, and G. Bersuker, "A Novel Bias Temperature Instability Characterization Methodology for High-k nMOSFETs," *IEEE Transactions on Electron Devices*, vol. 27, no. 10, pp. 849–851, October 2006.
- [24] Z.Y. Liu, D.M. Huang, W.J. Liu, C.C. Liao, L.F. Zhang, , Z.H. Gan, W.S. Wong, and M.-F. Li, "Comprehensive Studies of BTI Degradation in SiON Gate Dielectric CMOS Transistors by New Measurement Techniques," in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2008.

- [25] M.-F. Li, D.M. Huang, C. Shen, T. Yang, W.J. Liu, and Z.Y. Liu, "Understand NBTI Mechanism by Developing Novel Measurement Techniques," *IEEE Transactions on Device and Materials Reliability*, vol. 8, no. 1, pp. 62–71, March 2008.
- [26] G. A. Du, D. S. Ang, Y. Z. Hu, S. Wang, and C. M. Ng, "Physical Framework for NBTI: Insight from Ultra-Fast Switching Measurement of NBTI Recovery," in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2008, pp. 735–736.
- [27] C. Shen, M.F. Li, X.P. Wang, H.Y. Yu, Y. P. Feng, A. T.-L. Lim, Y.C. Yeo, D.S.H. Chan, and D.L. Kwong, "Negative U Traps in HfO₂ Gate Dielectrics and Frequency Dependence of Dynamic BTI in MOSFETs," in *Proceedings of IEEE International Electron Devices Meeting (IEDM)*, 2004, pp. 733–736.
- [28] M. Denais, A. Bravaix, V. Huard, C. Parthasarathy, G. Ribes, F. Perrier, Y. Rey Tauriac, and N.Revil, "On-the-Fly Characterization of NBTI in Ultra-Thin Gate Oxide PMOSFETs," in *Proceedings of IEEE International Electron Devices Meeting (IEDM)*, 2004.
- [29] T. Grasser and B. Kaczer, "Negative Bias Temperature Instability: Recoverable versus Permanent Degradation," in *Proceedings of IEEE European Solid-State Device Research Conference (ESSDERC)*, 2007, pp. 127–130.
- [30] T. Grasser, B. Kaczer, Ph. Hehenberger, W. Gös, R. O'Connor, H. Reisinger, W. Gustin, and C. Schlünder, "Simultaneous Extraction of Recoverable and Permanent Components Contributing to Bias-Temperature Instability," in *Proceedings of IEEE International Electron Devices Meeting (IEDM)*, 2007.
- [31] M. A. Alam, H. Kufuoglu, D. Varghese, and S. Mahapatra, "A Comprehensive Model for PMOS NBTI Degradation: Recent Progress," *Microelectronics Reliability*, vol. 47, pp. 853–862, 2007.
- [32] D. Varghese, S. Mahapatra, and M. A. Alam, "Hole Energy Dependent Interface Trap Generation in MOSFET Si/SiO₂ Interface," *IEEE Electron Device Letters*, vol. 26, no. 8, pp. 572–574, 2005.
- [33] A.T. Krishnan, C. Chancellor, S. Chakravarthi, P.E. Nicollian, V. Reddy, A. Varghese, R.B. Khamankar, and S. Krishnan, "Material Dependence of Hydrogen Diffusion: Implications for NBTI Degradation," in *Proceedings of IEEE International Electron Devices Meeting (IEDM)*, 2005, pp. 688–691.
- [34] J. F. Zhang and M. H. Chang, "An Assessment of Effective Mobility Variation during Negative Bias Temperature Instability," *Electrochemical Society Transactions (ECS)*, vol. 6, no. 3, pp. 301–311, 2007.
- [35] M. A. Alam, "A Review of New Characterization Methodologies of Gate Dielectric Breakdown and Negative Bias Temperature Instability," in *Proceedings of IEEE International Symposium on Physical and Failure Analysis of Integrated Circuits (IPFA)*, 2006, pp. 25–32.
- [36] A.T. Krishnan, V. Reddy, S. Chakravarthi, J. Rodriguez, S. John, and S. Krishnan, "NBTI Impact on Transistor & Circuit: Models, Mechanisms & Scaling Effects," in *Proceedings of IEEE International Electron Devices Meeting (IEDM)*, 2003, pp. 1–4.

- [37] Y. Tsididis, *Operation and Modeling of the MOS Transistor*, Oxford University Press, Inc., 198 Madison Avenue, New York 10016, 2nd edition, 1999, ISBN 0-19-517014-8.
- [38] S. E. Rauch, “The Statistics of NBTI-Induced V_T and β Mismatch Shifts in pMOSFETs,” *IEEE Transactions on Device and Materials Reliability*, vol. 2, no. 4, pp. 89–93, 2002.
- [39] T. Grasser, P.-J. Wagner, Ph. Hehenberger, W. Gös, and B. Kaczer, “A Rigorous Study of Measurement Techniques for Negative Bias Temperature Instability,” *IEEE Transactions on Device and Materials Reliability*, vol. 8, no. 3, pp. 526 – 535, 2008.
- [40] C. Shen, M.-F. Li, C.E. Foo, T. Yang, D.M. Huang, A. Yap, G.S. Samudra, and Y.-C. Yeo, “Characterization and Physical Origin of Fast V_{th} Transient in NBTI of pMOSFETs with SiON Dielectric,” in *Proceedings of IEEE International Electron Devices Meeting (IEDM)*, 2006, pp. 333–336.
- [41] A.E. Islam, E.N. Kumar, H. Das, S. Purawat, V. Maheta, H. Aono, E. Murakami, S. Mahapatra, and M.A. Alam, “Theory and Practice of On-The-Fly and Ultra-Fast VT Measurements for NBTI Degradation: Challenges and Opportunities,” in *Proceedings of IEEE International Electron Devices Meeting (IEDM)*, 2007, pp. 805–808.
- [42] J.F. Zhang, Z. Ji, M.H. Chang, B. Kaczer, and G. Groeseneken, “Real V_{th} Instability of pMOSFETs under Practical Operation Conditions,” in *Proceedings of IEEE International Electron Devices Meeting (IEDM)*, 2007, pp. 817–820.
- [43] J.S. Brugler and P. Jespers, “Charge Pumping in MOS Devices,” *IEEE Transactions on Electron Devices*, vol. 16, pp. 297–302, 1969.
- [44] G. Groeseneken, H.E. Maes, N. Beltran, and R. F. de Keersmaecker, “A Reliable Approach to Charge-Pumping Measurements in MOS Transistors,” *IEEE Transactions on Electron Devices*, vol. 31, pp. 42–53, 1984.
- [45] P. Heremans, J. Witters, G. Groeseneken, and H.E. Maes, “Analysis of the Charge Pumping Technique and its Application for the Evaluation of MOSFET Degradation,” *IEEE Transactions on Electron Devices*, vol. 36, pp. 1318–1335, 1989.
- [46] G.V.d. Bosch, G.V. Groeseneken, P. Heremans, and H.E. Maes, “Spectroscopic Charge Pumping: A New Procedure for Measuring Interface Trap Distributions on MOS Transistors,” *IEEE Transactions on Electron Devices*, vol. 38, pp. 1820–1831, 1991.
- [47] R.E. Paulsen and M.H. White, “Theory and Application of Charge Pumping for the Characterization of Si-SiO₂ Interface and Near-Interface Oxide Traps,” *IEEE Transactions on Electron Devices*, vol. 41, pp. 1213–1216, 1994.
- [48] D. Bauza, “Rigorous Analysis of Two-Level Charge Pumping: Application to the Extraction of Interface Trap Concentration versus Energy Profiles in Metal-Oxide-Semiconductor Transistors,” *Journal of Applied Physics*, vol. 94, pp. 3239–3248, 2003.
- [49] S. Mahapatra, K. Ahmed, D. Varghese, A. E. Islam, G. Gupta, L. Madhav, D. Saha, and M. A. Alam, “On the Physical Mechanism of NBTI in Silicon Oxynitride p-MOSFETs: Can Differences in Insulator Processing Conditions Resolve the Interface Trap Generation

- versus Hole Trapping Controversy?,” in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2007, pp. 1–9.
- [50] Th. Aichinger and M. Nelhiebel, “Charge Pumping Revisited - The Benefits of an Optimized Constant Base Level Charge Pumping Technique for MOS-FET Analysis,” in *Proceedings of IEEE International Integrated Reliability Workshop*, 2007.
- [51] W.J. Liu, Z.Y. Liu, D.M. Huang, C.C. Liao, L.F. Zhang, Z.H. Gan, W.S. Wong, C. Shen, and M.-F. Li, “On-The-Fly Interface Trap Measurement and its Impact on the Understanding of NBTI Mechanism for p-MOSFETs with SiON Gate Dielectric,” in *Proceedings of IEEE International Electron Devices Meeting (IEDM)*, 2007.
- [52] E.H. Nicollian and J.R. Brews, *MOS (Metal Oxide Semiconductor) Physics and Technology*, John Wiley & Sons - Interscience, 1982.
- [53] E.H. Nicollian and A. Goetzberger, “The Si-SiO₂ Interface – Electrical Properties as Determined by the Metal-Insulator-Silicon Conductance Technique,” *The Bell System Tech.J.*, vol. 46, no. 6, pp. 1055–1133, 1967.
- [54] J.W. McPherson, “Quantum Mechanical Treatment of Si-O Bond Breakage in Silica under Time Dependent Dielectric Breakdown Testing,” in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2007, pp. 209–216.
- [55] T. Grasser, W. Göss, and B. Kaczer, “Critical Modeling Issues in Negative Bias Temperature Instability,” in *Electrochemical Society Transactions (ECS)*, 2009.
- [56] S. Ogawa and N. Shiono, “Generalized Diffusion-Reaction Model for the Low-Field Charge-Buildup Instability at the Si-SiO₂ Interface,” *Physical Review B*, vol. 51, no. 7, pp. 4218–4230, 1995.
- [57] S. Ogawa, M. Shimaya, and N. Shiono, “Interface-Trap Generation at Ultrathin SiO₂ (4-6 nm)-Si Interfaces During Negative-Bias Temperature Aging,” *Journal of Applied Physics*, vol. 77, no. 3, pp. 1137–1148, 1995.
- [58] A.T. Krishnan, S. Chakravarthi, P. Nicollian, V. Reddy, and S. Krishnan, “Negative Bias Temperature Instability Mechanism: The Role of Molecular Hydrogen,” *Applied Physics Letters*, vol. 88, no. 15, pp. 1–3, 2006.
- [59] M.A. Alam and S. Mahapatra, “A Comprehensive Model of PMOS NBTI Degradation,” *Microelectronics Reliability*, vol. 45, pp. 71–81, 2005.
- [60] S. Chakravarthi, A.T. Krishnan, V. Reddy, C.F. Machala, and S. Krishnan, “A Comprehensive Framework For Predictive Modeling of Negative Bias Temperature Instability,” in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2004, pp. 273–282.
- [61] T. Grasser, W. Göss, V. Sverdlov, and B. Kaczer, “The Universality of NBTI Relaxation and its Implications for Modeling and Characterization,” in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2007, pp. 268–280.

- [62] V. Huard, C. Parthasarathy, N. Rallet, C. Guerin, M. Mammase, D. Barge, and C. Ouvrard, “New Characterization and Modeling Approach for NBTI Degradation from Transistor to Product Level,” in *Proceedings of IEEE International Electron Devices Meeting (IEDM)*, 2007, pp. 797–800.
- [63] M. A. Alam and H. Kufluoglu, “On Quasi-Saturation of Negative Bias Temperature Degradation,” in *Electrochemical Society Transactions (ECS)*, 2005, pp. 139–145.
- [64] H. Küfflüoglu and M.A. Alam, “A Generalized Reaction–Diffusion Model with Explicit H–H₂ Dynamics for Negative-Bias-Temperature-Instability (NBTI) Degradation,” *IEEE Transactions on Electron Devices*, vol. 54, no. 5, pp. 1101–1107, 2007.
- [65] M. Houssa, M. Aoulaiche, S. De Gendt, G. Groeseneken, M. M. Heyns, and A. Stesmans, “Reaction-Dispersive Proton Transport Model for Negative Bias Temperature Instabilities,” *Applied Physics Letters*, vol. 86, pp. 1–3, 2005.
- [66] S. Zafar, “Statistical Mechanics Based Model for Negative Bias Temperature Instability Induced Degradation,” *Journal of Applied Physics*, vol. 97, pp. 1–9, 2005.
- [67] B. Kaczer, V. Arkhipov, R. Degraeve, N. Collaert, G. Groeseneken, and M. Goodwin, “Temperature Dependence of the Negative Bias Temperature Instability in the Framework of Dispersive Transport,” *Applied Physics Letters*, vol. 85, pp. 1–3, 2005.
- [68] A. Haggag, G. Anderson, S. Parihar, D. Burnett, G. Abeln, J. Higman, and M. Moosa, “Understanding SRAM High-Temperature-Operating-Life NBTI: Statistics and Permanent vs. Recoverable Damage,” in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2007, pp. 452–456.
- [69] D. S. Ang, S. Wang, G. A. Du, and Y. Z. Hu, “A Consistent Deep-Level Hole Trapping Model for Negative Bias Temperature Instability,” *IEEE Transactions on Device and Materials Reliability*, vol. 8, no. 1, pp. 22–34, 2008.
- [70] A. Stesmans, “Dissociation Kinetics of Hydrogen-Passivated P_b Defects at the (111) Si/SiO₂ Interface,” *Physical Review B*, vol. 61, no. 12, pp. 8393–8403, 2000.
- [71] T. Grasser, W. Göss, and B. Kaczer, “Dispersive Transport and Negative Bias Temperature Instability: Boundary Conditions, Initial Conditions, and Transport Models,” *IEEE Transactions on Device and Materials Reliability*, vol. 8, no. 1, pp. 79–97, 2008.
- [72] A. Haggag, W. McMahon, K. Hess, K. Cheng, J. Lee, and J. Lyding, “High-Performance Chip Reliability from Short-Time-Tests,” in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2001, pp. 271–279.
- [73] T. Aichinger, M. Nelhiebel, and T. Grasser, “Unambiguous Identification of the NBTI Recovery Mechanism using Ultra-Fast Temperature Changes,” in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2009.
- [74] D.S. Ang, Z.Q. Teo, and C.M. Ng, “Reassessing NBTI Mechanisms by Ultrafast Charge Pumping Measurement,” in *Proceedings of IEEE International Integrated Reliability Workshop*, 2009.

- [75] B. Kaczer, T. Grasser, R. Fernandez, and G. Groeseneken, "Toward Understanding the Wide Distribution of Time Scales in Negative Bias Temperature Instability," in *Electrochemical Society Transactions (ECS)*, 2007, pp. 265–281.
- [76] W.J. Liu, Z.Y. Liu, D.M. Huang, Y. Luo, C.C. Liao, L.F. Zhang, Z.H. Gan, W. Wong, and M.-F. Li, "Investigations of NBTI by Conventional and New Measurement Methods for p-MOSFETs," in *Proceedings of IEEE International Nanoelectronics Conference (INEC)*, 2008.
- [77] T. Grasser, B. Kaczer, and W. Gös, "An Energy-Level Perspective of Bias Temperature Instability," in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2008, 28-38.
- [78] Ph. Hehenberger, Th. Aichinger, T. Grasser, W. Gös, O. Triebel, B. Kaczer, and M. Nelhiebel, "Do NBTI-Induced Interface States Show Fast Recovery? A Study Using a Corrected On-The-Fly Charge-Pumping Measurement Technique," in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2009, pp. 1033–1038.
- [79] S. Mahapatra, A.E. Islam, S. Deora, V.D. Maheta, K. Joshi, A. Jain, and M.A. Alam, "A Critical Re-evaluation of the Usefulness of R-D Framework in Predicting NBTI Stress and Recovery," in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, april 2011, pp. 6A.3.1 –6A.3.10.
- [80] "JEDEC, Global Standards for the Microelectronics Industry," <http://www.jedec.org/events-meetings>.
- [81] M.F. Li, G. Chen, C. Shen, X.P. Wang, H.Y. Yu, Y.-C. Yeo, and D.L. Kwong, "Dynamic Bias-Temperature Instability in Ultrathin SiO₂ and HfO₂ Metal-Oxide-Semiconductor Field Effect Transistors and Its Impact on Device Lifetime," *Japanese Journal of Applied Physics*, vol. 43, no. 11B, pp. 7807–7814, 2004.
- [82] M. Denais, A. Bravaix, V. Huard, C. Parthasarathy, C. Guerin, G. Ribes, F. Perrier, M. Mairy, and D. Roy, "Paradigm Shift for NBTI Characterization in Ultra-Scaled CMOS Technologies," in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2006, pp. 735–736.
- [83] J. Kakalios, R.A. Street, and W.B. Jackson, "Stretched-Exponential Relaxation Arising from Dispersive Diffusion of Hydrogen in Amorphous Silicon," *Physical Review Letters*, vol. 59, no. 9, pp. 1037–1040, 1987.
- [84] V. Huard, C.R. Parthasarathy, C. Guerin, and M. Denais, "Physical Modeling of Negative Bias Temperature Instabilities for Predictive Extrapolation," in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2006.
- [85] M. A. Alam, "A Critical Examination of the Mechanics of Dynamic NBTI for PMOSFETs," in *Proceedings of IEEE International Electron Devices Meeting (IEDM)*, 2003, pp. 345–348.
- [86] S. Tsujikawa, T. Mine, K. Watanabe, Y. Shimamoto, R. Tsuchiya, K. Ohnishi, T. Onai, J. Yugami, and S. Kimura, "Negative Bias Temperature Instability of pMOSFETs with Ultra-Thin SiON Gate Dielectrics," in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2003, pp. 183–188.

- [87] S. Tsujikawa and J. Yugami, "Positive Charge Generation due to Species of Hydrogen during NBTI Phenomenon in pMOSFETs with Ultra-Thin SiON Gate Dielectrics," *Microelectronics Reliability*, vol. 45, no. 1, pp. 65–69, 2005.
- [88] J.F. Zhang, C.Z. Zhao, A.H. Chen, G. Groeseneken, and R. Degraeve, "Hole Traps in Silicon Dioxides — Part I: Properties," *IEEE Transactions on Electron Devices*, vol. 51, no. 8, pp. 1267–1273, 2004.
- [89] I μ E, *MINIMOS-NT 2.1 User's Guide*, Institut für Mikroelektronik, Technische Universität Wien, Austria, 2004, <http://www.iue.tuwien.ac.at/software/minimos-nt>.
- [90] C. Jungemann, T. Grasser, B. Neinhüus, and B. Meinerzhagen, "Failure of Moments-Based Transport Models in Nanoscale Devices Near Equilibrium," *IEEE Transactions on Electron Devices*, vol. 52, no. 11, pp. 2404–2408, November 2005.
- [91] J.P. Campbell, P.M. Lenahan, A.T. Krishnan, and S. Krishnan, "Location, Structure, and Density of States of NBTI-Induced Defects in Plasma Nitrided pMOSFETS," in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2007, pp. 503–510.
- [92] V.M. Agostinelli, H. Shin, and A.F. Tasch, "A Comprehensive Model for Inversion Layer Hole Mobility for Simulation of Submicrometer MOSFET's," *IEEE Transactions on Electron Devices*, vol. 38, no. 1, pp. 151–159, January 1991.
- [93] L.B. Freeman and W.E. Dahlke, "Theory of tunneling into interface states," *Solid-State Electron.*, vol. 13, no. 11, pp. 1483–1503, 1970.
- [94] M. Denais, V. Huard, C. Parthasarathy, G. Ribes, F. Perrier, N. Revil, and A. Bravaix, "Interface Trap Generation and Hole Trapping under NBTI and PBTI in Advanced CMOS Technology with a 2-nm Gate Oxide," *IEEE Transactions on Device and Materials Reliability*, vol. 4, pp. 715–722, 2004.
- [95] P. Habaš, *Analysis of Physical Effects in Small Silicon MOS Devices*, Ph.D. thesis, Technical University of Vienna, 1993.
- [96] Th. Aichinger, "Implementation of the Charge Pumping Method for MOS Characterization into Existing Soft- and Hardware Laboratory Environment," M.S. thesis, Technical University of Graz, 2007.
- [97] Th. Aichinger, M. Nelhiebel, and T. Grasser, "On the Energy Dependence of Oxide Trap Recovery after NBTI Stress," in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2009.
- [98] T. Grasser, B. Kaczer, W. Gös, Th. Aichinger, Ph. Hehenberger, and M. Nelhiebel, "A Two-Stage Model for Negative Bias Temperature Instability," in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2009, pp. 33–44.
- [99] D.M. Fleetwood, "Fast and Slow Border Traps in MOS Devices," in *Proceedings of Radiation and its Effects on Components and Systems (RADECS)*, 1995.

- [100] B. Kaczer, T. Grasser, J. Martin Martinez, E. Simoen, M. Aoulaiche, Ph. J. Roussel, and G. Groeseneken, "NBTI from the Perspective of Defect States with Widely Distributed Time Scales," in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2009, pp. 55–60.
- [101] W.J. Liu, D.M. Huang, Q.Q. Sun, C.C. Liao, L.F. Zhang, Z.H. Gan, W. Wong, and M.-F. Li, "Studies of NBTI in pMOSFETs with Thermal and Plasma Nitrided SiON Gate Oxides by OFIT and FPM Methods," in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2009.
- [102] E.N. Kumar, V.D. Maheta, S. Purawat, A.E. Islam, C. Olsen, K. Ahmed, M.A. Alam, and S. Mahapatra, "Material Dependence of NBTI Physical Mechanism in Silicon Oxynitride (SiON) pMOSFETs: A Comprehensive Study by Ultra-Fast On-The-Fly (UF-OTF) I_D LIN Technique," in *Proceedings of IEEE International Electron Devices Meeting (IEDM)*, 2007, pp. 809–812.
- [103] S. Mahapatra and M.A. Alam, "Defect Generation in p-MOSFETs Under Negative-Bias Stress: An Experimental Perspective," *IEEE Transactions on Device and Materials Reliability*, vol. 8, no. 1, pp. 35–46, March 2008.
- [104] S. Mahapatra, V.D. Maheta, A.E. Islam, and M.A. Alam, "Isolation of NBTI Stress Generated Interface Trap and Hole-Trapping Components in PNO p-MOSFETs," *IEEE Transactions on Electron Devices*, vol. 56, no. 2, pp. 236–242, Feb. 2009.
- [105] R.G. Southwick, W.B. Knowlton, B. Kaczer, and T. Grasser, "On the Thermal Activation of Negative Bias Temperature Instability," in *Proceedings of IEEE International Integrated Reliability Workshop*, 2009.
- [106] T. Grasser and B. Kaczer, "Evidence that Two Tightly Coupled Mechanisms are Responsible for Negative Bias Temperature Instability in Oxynitride MOSFETs," *IEEE Transactions on Electron Devices*, vol. 56, no. 5, pp. 1056–62, 2009.
- [107] V.D. Maheta, E.N. Kumar, S. Purawat, C. Olsen, K. Ahmed, and S. Mahapatra, "Development of an Ultrafast On-the-Fly I_D LIN Technique to Study NBTI in Plasma and Thermal Oxynitride p-MOSFETs," *IEEE Transactions on Electron Devices*, vol. 55, no. 10, pp. 2614–2622, Oct. 2008.
- [108] T. Grasser, P.-J. Wagner, Ph. Hehenberger, W. Gös, and B. Kaczer, "A Rigorous Study of Measurement Techniques for Negative Bias Temperature Instability," in *Proceedings of IEEE International Integrated Reliability Workshop*, 2007, pp. 6–11.
- [109] C. Schlünder, R.-P. Vollertsen, W. Gustin, and H. Reisinger, "A Reliable and Accurate Approach to Assess NBTI Behavior of State-of-the-Art pMOSFETs with Fast-WLR," in *Proceedings of IEEE European Solid-State Device Research Conference (ESSDERC)*, 2007, pp. 131–134.
- [110] Ph. Hehenberger, P.-J. Wagner, H. Reisinger, and T. Grasser, "Comparison of Fast Measurement Methods for Short-Term Negative Bias Temperature Stress and Relaxation," in *Proceedings of IEEE European Solid-State Device Research Conference (ESSDERC)*, 2009, pp. 311–314.

- [111] H. Reisinger, T. Grasser, and C. Schlünder, “A Study of NBTI by the Statistical Analysis of the Properties of Individual Defects in pMOSFETs,” in *Proceedings of IEEE International Integrated Reliability Workshop*, 2009, pp. 30–35.
- [112] T. Grasser, H. Reisinger, P.-J. Wagner, F. Schanovsky, W. Gös, and B. Kaczer, “The Time Dependent Defect Spectroscopy (TDDS) for the Characterization of the Bias Temperature Instability,” in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2010, pp. 16–25.
- [113] V. Huard, C.R. Parthasarathy, and M. Denais, “Single-Hole Detrapping Events in pMOSFETs NBTI Degradation,” in *Proceedings of IEEE International Integrated Reliability Workshop*, 2005, p. 5.
- [114] B. Kaczer, T. Grasser, Ph. J. Roussel, J. Franco, R. Degraeve, L.-A. Ragnarsson, E. Simoen, G. Groeseneken, and H. Reisinger, “Origin of NBTI Variability in Deeply Scaled pFETs,” in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2010.
- [115] T. Grasser, H. Reisinger, W. W. Gös, Th. Aichinger, Ph. Hehenberger, P.-J. Wagner, M. Nelhiebel, J. Franco, and B. Kaczer, “Switching Oxide Traps as the Missing Link Between Negative Bias Temperature Instability and Random Telegraph Noise,” in *Proceedings of IEEE International Electron Devices Meeting (IEDM)*, 2009.
- [116] H. Reisinger, T. Grasser, W. Gustin, and C. Schlünder, “The Statistical Analysis of Individual Defects Constituting NBTI and its Implications for Modeling DC- and AC-Stress,” in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2010, pp. 7–15.
- [117] D.V. Lang, “Deep-Level Transient Spectroscopy: A New Method to Characterize Traps in Semiconductors,” *Journal of Applied Physics*, vol. 45, no. 7, pp. 3023–3032, 1974.
- [118] A. Karwath and M. Schulz, “Deep Level Transient Spectroscopy on Single, Isolated Interface Traps in Field-Effect Transistors,” *Applied Physics Letters*, vol. 52, no. 8, pp. 634–636, 1988.
- [119] T. Grasser, “Negative Bias Temperature Instability: Modeling Challenges and Perspectives,” in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2008.
- [120] A.E. Islam, H. Kufluoglu, D. Varghese, S. Mahapatra, and M.A. Alam, “Recent Issues in Negative-Bias Temperature Instability: Initial Degradation, Field Dependence of Interface Trap Generation, Hole Trapping Effects, and Relaxation,” *IEEE Transactions on Electron Devices*, vol. 54, no. 9, pp. 2143–2154, 2007.
- [121] A.L. McWhorter, “ $1/f$ Noise and Germanium Surface Properties,” *Sem.Surf.Phys University of Pennsylvania Press*, pp. 207–228, 1957.
- [122] W. Shockley and W.T. Read, “Statistics of the Recombinations of Holes and Electrons,” *Physical Review*, vol. 87, no. 5, pp. 835–842, 1952.
- [123] M. Masuduzzaman, A.E. Islam, and M.A. Alam, “Exploring the Capability of Multifrequency Charge Pumping in Resolving Location and Energy Levels of Traps Within Dielectric,” *IEEE Transactions on Electron Devices*, vol. 55, no. 12, pp. 3421–3431, 2008.

- [124] M.J. Kirton and M.J. Uren, “Noise in Solid-State Microstructures: A New Perspective on Individual Defects, Interface States and Low-Frequency ($1/f$) Noise,” *Advances in Physics*, vol. 38, pp. 367–468, 1989.
- [125] C.H. Henry and D.V. Lang, “Nonradiative Capture and Recombination by Multiphonon Emission in GaAs and GaP,” *Physical Review B*, vol. 15, no. 2, pp. 989–1016, 1977.
- [126] S. Makram Ebeid and M. Lannoo, “Quantum model for phonon-assisted tunnel ionization of deep levels in a semiconductor,” *Physical Review B*, vol. 25, pp. 6406–6424, 1982.
- [127] S. Makram Ebeid and M. Lannoo, “Electric-Field-Induced Phonon-Assisted Tunnel Ionization from Deep Levels in Semiconductors,” *Physical Review Letters*, vol. 48, no. 18, pp. 1281–1284, 1982.
- [128] S.D. Ganichev, I.N. Yassievich, V.I. Perel, H. Ketterl, and W. Prettl, “Tunnelling ionization of deep centres in high-frequency electric fields,” *J.Phys.:Condensed Matter*, vol. 14, pp. 1263–1295, 2002.
- [129] A. Gehring, T. Grasser, H. Kosina, and S. Selberherr, “Simulation of Hot-Electron Oxide Tunneling Current Based on a Non-Maxwellian Electron Energy Distribution Function,” *Journal of Applied Physics*, vol. 92, no. 10, pp. 6019–6027, 2002.
- [130] T. Grasser, “Stochastic Charge Trapping in Oxides: From Random Telegraph Noise to Bias Temperature Instabilities,” *Microelectronics Reliability*, vol. XX, pp. 1–33, 2011.
- [131] M. Schulz and N.M. Johnson, “Transient Capacitance Measurements of Hole Emission from Interface States in MOS Structures,” *Applied Physics Letters*, vol. 31, no. 9, pp. 622–625, 1977.
- [132] T. L. Tewksbury, *Relaxation Effects in MOS Devices due to Tunnel Exchange with Near-Interface Oxide Trap*, Ph.D. thesis, MIT, 1992.
- [133] F. Jimenez Molinos, A. Palma, F. Gamiz, J. Banqueri, and J.A. Lopez Villanueva, “Physical model for trap-assisted inelastic tunneling in metal-oxide-semiconductor structures,” *Journal of Applied Physics*, vol. 90, no. 7, pp. 3396–3404, 2001.
- [134] T. Grasser, B. Kaczer, T. Aichinger, W. Goes, and M. Nelhiebel, “Defect Creation Stimulated by Thermally Activated Hole Trapping as the Driving Force Behind Negative Bias Temperature Instability in SiO₂, SiON, and High-k Gate Stacks,” in *Proceedings of IEEE International Integrated Reliability Workshop*, 2008.
- [135] M.H. Woods and R. Williams, “Hole Trap in Silicon Dioxide,” *Journal of Applied Physics*, vol. 47, no. 3, pp. 1082–1089, 1976.
- [136] D.M. Fleetwood, H.D. Xiong, Z.-Y. Lu, C.J. Nicklaw, J.A. Felix, R.D. Schrimpf, and S.T. Pantelides, “Unified Model of Hole Trapping, $1/f$ Noise, and Thermally Stimulated Current in MOS Devices,” *IEEE Transactions on Nuclear Science*, vol. 49, no. 6, pp. 2674–2683, 2002.
- [137] P.E. Blöchl and J.H. Stathis, “Hydrogen Electrochemistry and Stress-Induced Leakage Current in Silica,” *Physical Review Letters*, vol. 83, no. 2, pp. 372–375, 1999.

- [138] P.E. Blöchl, “First-Principles Calculations of Defects in Oxygen-Deficient Silica Exposed to Hydrogen,” *Physical Review B*, vol. 62, no. 10, pp. 6158–6179, 2000.
- [139] P.M. Lenahan, “Atomic Scale Defects Involved in MOS Reliability Problems,” *Microelectronics Reliability*, vol. 69, pp. 173–181, 2003.
- [140] P.M. Lenahan and J.F. Conley, “A Comprehensive Physically Based Predictive Model for Radiation Damage in MOS Systems,” *IEEE Transactions on Nuclear Science*, vol. 45, no. 6, pp. 2413–2423, 1998.
- [141] C.J. Nicklaw, Z.-Y. Lu, D.M. Fleetwood, R.D. Schrimpf, and S.T. Pantelides, “The Structure, Properties, and Dynamics of Oxygen Vacancies in Amorphous SiO₂,” *IEEE Transactions on Nuclear Science*, vol. 49, no. 6, pp. 2667–2673, 2002.
- [142] P.-J. Wagner, T. Aichinger, T. Grasser, M. Nelhiebel, and L.K.J. Vandamme, “Possible Correlation between Flicker Noise and Bias Temperature Stress,” in *Proceedings of the 20th International Conference on Noise and Fluctuations*, 2009, pp. 621–624.
- [143] M.B. Weissman, “1/f Noise and Other Slow, Nonexponential Kinetics in Condensed Matter,” *Review of Modern Physics*, vol. 60, no. 2, pp. 537–571, 1988.
- [144] D.M. Fleetwood, P.S. Winokur, Jr. R.A. Reber, T.L. Meisenheimer, J.R. Schwank, M.R. Shaneyfelt, and L.C. Riewe, “Effects of Oxide Traps, Interface Traps, and “Border Traps” on Metal-Oxide-Semiconductor Devices,” *Journal of Applied Physics*, vol. 73, pp. 5058–5074, 1993.
- [145] S.D. Ganichev and W. Prettl, “Deep Impurity-Center Ionization by Far-Infrared Radiation,” *Phys.Solid State*, vol. 39, no. 11, pp. 1703–1726, 1997.
- [146] D. Ielmini, M. Manigrasso, F. Gattel, and M.G. Valentini, “A New NBTI Model Based on Hole Trapping and Structural Relaxation in MOS Dielectrics,” *IEEE Transactions on Electron Devices*, vol. 56, no. 9, pp. 1943–1952, 2009.
- [147] A.J. Lelis and T.R. Oldham, “Time Dependence of Switching Oxide Traps,” *IEEE Transactions on Nuclear Science*, vol. 41, no. 6, pp. 1835–1843, 1994.
- [148] E.H. Poindexter and W.L. Warren, “Paramagnetic Point Defects in Amorphous Thin Films of SiO₂ and Si₃N₄: Updates and Additions,” *J.Electrochem.Soc.*, vol. 142, no. 7, pp. 2508–2516, 1995.
- [149] F. Schanovsky, W. Gös, and T. Grasser, “Ab-Initio Calculation of the Vibrational Influence on Hole-Trapping,” in *Proc. Intl. Workshop Comp.Electronics (IWCE)*, 2010.
- [150] F. Schanovsky, W. Gös, and T. Grasser, “An Advanced Description of Oxide Traps in MOS Transistors and its Relation to DFT,” *Journal of Computational Electronics*, vol. 9, no. 4, pp. 135–140, 2010.
- [151] F. Schanovsky, W. Gös, and T. Grasser, “Multiphonon Hole Trapping from First Principles,” *Journal of Vacuum Science and Technology B*, vol. 29, no. 1, pp. 01A201–1 – 01A201–5, 2011.
- [152] J. Franck, “Elementary Processes of Photochemical Reactions,” *Transactions of the Faraday Society*, vol. 21, pp. 536–542, 1926.

- [153] K. Huang and A. Rhys, "Theory of Light Absorption and Non-Radiative Transitions in F-Centres," *Proceedings of the Royal Society A*, vol. 204, pp. 406–423, 1950.
- [154] J.K. Rudra and W.B. Fowler, "Oxygen Vacancy and the E'_1 Center in Crystalline SiO_2 ," *Physical Review B*, vol. 35, no. 15, pp. 8223–8230, 1987.
- [155] S.P. Karna, A.C. Pineda, R.D. Pugh, W.M. Shedd, and T.R. Oldham, "Electronic Structure Theory and Mechanisms of the Oxide Trapped Hole Annealing Process," *IEEE Transactions on Nuclear Science*, vol. 47, no. 6, pp. 2316–2321, 2000.
- [156] T.H. Keil, "Shapes of Impurity Absorption Bands in Solids," *Physical Review*, vol. 140, no. 2A, pp. A601–A617, 1965.
- [157] W.B. Fowler, J.K. Rudra, M.E. Zvanut, and F.J. Feigl, "Hysteresis and Franck-Condon Relaxation in Insulator-Semiconductor Tunneling," *Physical Review B*, vol. 41, no. 12, pp. 8313–8317, 1990.
- [158] V.N. Abakumov, V.I. Perel, and I.N. Yassievich, *Nonradiative Recombination in Semiconductors*, vol. 33, North-Holland, 1991.
- [159] A.M. Stoneham, "Non-radiative transitions in semiconductors," *Rep.Prog.Phys.*, vol. 44, pp. 1255, 1981.
- [160] W.B. Fowler, *Physics of Color Centers - Electronic States and Optical Transitions of Color Centers*, Academic Press, New York, 1968.
- [161] W. Gös, F. Schanovsky, Ph. Hehenberger, P.J. Wagner, and T. Grasser, "Charge Trapping and the Negative Bias Temperature Instability," in *Electrochemical Society Transactions (ECS)*, S. Kar, M. Houssa, S. Van Elshocht, D. Landheer, D. Misra, and K. Kita, Eds., 2010, pp. 565–589.
- [162] T. Grasser, B. Kaczer, W. Gös, H. Reisinger, T. Aichinger, Ph. Hehenberger, P.-J. Wagner, F. Schanovsky, J. Franco, M.T. Luque, and M. Nelhiebel, "The Paradigm Shift in Understanding the Bias Temperature Instability: From Reaction-Diffusion to Switching Oxide Traps," *IEEE Transactions on Electron Devices*, vol. XX, pp. 1–15, 2011.
- [163] C.S. Kelley, "Moments of Semiclassical and Classical Absorption and Emission Band Shapes of Impurities in Solids," *Physical Review B*, vol. 20, no. 12, pp. 5084–5089, 1979.
- [164] J. Franco, B. Kaczer, G. Eneman, J. Mitard, A. Stesmans, V. Afanas'ev, T. Kauerauf, Ph. J. Roussel, M. Toledano Luque, M. Cho, R. Degraeve, T. Grasser, L.-Å. Ragnarsson, L. Witters, J. Tseng, S. Takeoka, W.-E. Wang, T. Y. Hoffmann, and G. Groeseneken, "6Å EOT $\text{Si}_{0.45}\text{Ge}_{0.55}$ pMOSFET with Optimized Reliability ($V_{\text{DD}} = 1\text{V}$): Meeting the NBTI Lifetime Target at Ultra-Thin EOT," in *Proceedings of IEEE International Electron Devices Meeting (IEDM)*, 2010, pp. 70–73.
- [165] J. Franco, B. Kaczer, M. Cho, G. Eneman, G. Groeseneken, and T. Grasser, "Improvements of NBTI Reliability in SiGe p-FETs," in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2010, pp. 1082–1085.

- [166] M.M. Rieger and P. Vogl, “Electron-Band Parameters in Strained $\text{Si}_{1-x}\text{Ge}_x$ Alloys on $\text{Si}_{1-y}\text{Ge}_y$ Substrates,” *Physical Review B*, vol. 48, no. 19, pp. 14276–14287, 1993.
- [167] M.V. Fischetti and S.E. Laux, “Band Structure, Deformation Potentials, and Carrier Mobility in Strained Si, Ge, and SiGe Alloys,” *Journal of Applied Physics*, vol. 80, no. 4, pp. 2234–2252, 1996.
- [168] “Ioffe Physical Technical Institute,” <http://www.ioffe.ru/SVA/NSM/Semicond/SiGe/bandstr.html>.
- [169] T. Grasser, “Charge Trapping in Oxides from RTN to BTI,” in *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, 2011.
- [170] M.J. Uren, M.J. Kirton, and S. Collins, “Anomalous Telegraph Noise in Small-Area Silicon Metal-Oxide-Semiconductor Field-Effect-Transistors,” *Physical Review B*, vol. 37, pp. 8346–8350, 1988.
- [171] M. Karner, A. Gehring, S. Holzer, M. Pourfath, M. Wagner, W. Gös, M. Vasicek, O. Baumgartner, C. Kernstock, K. Schnass, G. Zeiler, T. Grasser, H. Kosina, and S. Selberherr, “A Multi-Purpose Schrödinger-Poisson Solver for TCAD Applications,” *Journal of Computational Electronics*, vol. 6, no. 1-3, pp. 179–182, 2007.
- [172] T. Grasser and S. Selberherr, “Modeling of Negative Bias Temperature Instability,” *Journal of Telecommunications and Information Technology*, vol. 7, no. 2, pp. 92–102, 2007.
- [173] J. Crank, *The Mathematics of Diffusion*, 0-19-853411-6. Clarendon Press, second edition, 1975.

Own Publications

- [1] T. Grasser, B. Kaczer, W. Gös, H. Reisinger, T. Aichinger, Ph. Hehenberger, P.-J. Wagner, F. Schanovsky, J. Franco, M. Toledano-Luque, and M. Nelhiebel, “The Paradigm Shift in Understanding the Bias Temperature Instability: From Reaction-Diffusion to Switching Oxide Traps,” *IEEE Transactions on Electron Devices*, vol. XX, 2011, Invited Journal.
- [2] Ph. Hehenberger, W. Gös, O. Baumgartner, J. Franco, B. Kaczer, and T. Grasser, “Quantum-Mechanical Modeling of NBTI in High-k SiGe MOSFETs,” in *Proceedings of the 15th International Conference on Simulation of Semiconductor Processes and Devices*, 2011, pp. 11–14, Talk at the SISPAD, Osaka, Japan; 2011-09-08 – 2011-09-10.
- [3] J. Franco, B. Kaczer, M. Toledano-Luque, Ph. J. Roussel, Ph. Hehenberger, T. Grasser, J. Mitard, G. Eneman, T.Y. Hoffmann, and G. Groeseneken, “On the Impact of the Si Passivation Layer Thickness on the NBTI of Nanoscaled Si_{0.45}Ge_{0.55} pMOSFETs,” *Microelectronic Engineering*, vol. 88, no. 7, pp. 1388–1391, 2011, Journal.
- [4] T. Grasser, B. Kaczer, W. Gös, H. Reisinger, T. Aichinger, Ph. Hehenberger, P.-J. Wagner, F. Schanovsky, J. Franco, Ph. J. Roussel, and M. Nelhiebel, “Recent Advances in Understanding the Bias Temperature Instability,” in *Proceedings of the 2010 IEEE International Electron Devices Meeting (IEDM)*, 2010, pp. 82–85, Invited Talk at the IEDM, San Francisco; 2010-12-06 – 2010-12-08.
- [5] W. Gös, F. Schanovsky, Ph. Hehenberger, P.-J. Wagner, and T. Grasser, “Charge Trapping and the Negative Bias Temperature Instability,” in *Physics and Technology of High-k Materials 8*, pp. 565–589. ECS Transactions, 2010, Invited Book Contribution.
- [6] W. Gös, F. Schanovsky, Ph. Hehenberger, P.-J. Wagner, and T. Grasser, “Charge Trapping and the Negative Bias Temperature Instability,” in *Meet. Abstr. - Electrochem. Soc. 2010*, 2010, Talk at the 218th ECS Meeting, Las Vegas, USA; 2010-10-10 – 2010-10-15.
- [7] Ph. Hehenberger, H. Reisinger, and T. Grasser, “Recovery of Negative and Positive Bias Temperature Stress in pMOSFETs,” in *Final Report of IEEE International Integrated Reliability Workshop*, 2010, pp. 8–11, Talk at the IIRW, California, USA; 2010-10-17 – 2010-10-21.
- [8] T. Grasser, B. Kaczer, W. Gös, T. Aichinger, Ph. Hehenberger, and M. Nelhiebel, “Understanding Negative Bias Temperature Instability in the Context of Hole Trapping,” *Microelectronic Engineering*, vol. 86, no. 7-9, pp. 1876–1882, 2009, Invited Journal.
- [9] T. Grasser, H. Reisinger, W. Gös, T. Aichinger, Ph. Hehenberger, P.-J. Wagner, M. Nelhiebel, J. Franco, and B. Kaczer, “Switching Oxide Traps as the Missing Link Between Negative Bias

- Temperature Instability and Random Telegraph Noise,” in *Proceedings of the International Electron Devices Meeting*, 2009, Talk at the IEDM, Baltimore, USA; 2009-12-07 – 2009-12-09.
- [10] Ph. Hehenberger, P.-J. Wagner, H. Reisinger, and T. Grasser, “On the Temperature and Voltage Dependence of Short-Term Negative Bias Temperature Stress,” *Microelectronics Reliability*, vol. 49, pp. 1013–1017, 2009, Journal.
- [11] Ph. Hehenberger, P.-J. Wagner, H. Reisinger, and T. Grasser, “On the Temperature and Voltage Dependence of Short-Term Negative Bias Temperature Stress,” in *Proceedings of the 20th European Symposium on the Reliability of Electron Devices, Failure Physics and Analysis*, 2009, Talk at the ESREF, Bordeaux, France; 2009-10-05 – 2009-10-09.
- [12] Ph. Hehenberger, P.-J. Wagner, H. Reisinger, and T. Grasser, “Comparison of Fast Measurement Methods for Short-Term Negative Bias Temperature Stress and Relaxation,” in *Proceedings of the 39th European Solid-State Device Research Conference*, 2009, pp. 311–314, Talk at the ESSDERC, Athens, Greece; 2009-09-14 – 2009-09-18.
- [13] T. Grasser, B. Kaczer, W. Gös, T. Aichinger, Ph. Hehenberger, and M. Nelhiebel, “A Two-Stage Model for Negative Bias Temperature Instability,” in *2009 IEEE International Reliability Physics Symposium Proceedings*, 2009, pp. 33–44, Talk at the IRPS, Montreal, Canada; 2009-04-26 – 2009-04-30.
- [14] Ph. Hehenberger, T. Aichinger, T. Grasser, W. Gös, O. Triebel, B. Kaczer, and M. Nelhiebel, “Do NBTI-Induced Interface States Show Fast Recovery? A Study Using a Corrected On-The-Fly Charge-Pumping Measurement Technique,” in *2009 IEEE International Reliability Physics Symposium Proceedings*, 2009, pp. 1033–1038, Poster Presentation at the IRPS, Montreal, Canada; 2009-04-26 – 2009-04-30.
- [15] W. Gös, M. Karner, S. Tyaginov, Ph. Hehenberger, and T. Grasser, “Level Shifts and Gate Interfaces as Vital Ingredients in Modeling of Charge Trapping,” in *International Conference on Simulation of Semiconductor Processes and Devices 2008*, 2008, pp. 69–72, Talk at the SISPAD, Hakone, Japan; 2008-09-09 – 2008-09-11.
- [16] T. Grasser, P.-J. Wagner, Ph. Hehenberger, W. Gös, and B. Kaczer, “A Rigorous Study of Measurement Techniques for Negative Bias Temperature Instability,” *IEEE Transactions on Device and Materials Reliability*, vol. 8, no. 3, pp. 526–535, 2008, Journal.
- [17] T. Grasser, B. Kaczer, Ph. Hehenberger, W. Gös, R. Connor, H. Reisinger, W. Gustin, and C. Schlünder, “Simultaneous Extraction of Recoverable and Permanent Components Contributing to Bias-Temperature Instability,” in *International Electron Devices Meeting 2007*, 2007, pp. 801–804, Talk at the IEDM, Washington, DC, USA; 2007-12-10 – 2007-12-12.
- [18] T. Grasser, P.-J. Wagner, Ph. Hehenberger, W. Gös, and B. Kaczer, “A Rigorous Study of Measurement Techniques for Negative Bias Temperature Instability,” in *2007 IEEE International Integrated Reliability Workshop Final Report*, 2007, pp. 6–11, Talk at the IIRW, Fallen Leaf Lake, USA; 2007-10-15 – 2007-10-18.

Curriculum Vitae



Name: Philipp Paul Hehenberger
Date of Birth: October 12, 1980
Place of Birth: Vienna
Nationality: Austria
Marital status: married

- 1987 – 1991 Elementary school in Brunn am Gebirge
- 1991 – 1999 Secondary school at Kollegium Kalksburg in Vienna
with focus on languages and natural sciences
Final examination
- 1999 – 2000 Military service at the Federal Armed Forces of Austria
at the anti-aircraft defense academy staff division in Langenleobarn
- 2000 – 2006 Academic studies of technical physics
at the Vienna University of Technology
2003 Passed first diploma examination
- 2004 – 2005 Two summer internships at Infineon Technologies AG, Villach
2006 Diploma thesis in applied physics at Infineon Technologies AG, Villach
“Hot Carrier Stability of Trench Power MOSFETs under Avalanche Conditions”
Passed second diploma examination and received degree of “Diplom-Ingenieur”
- 2007 – 2012 PhD at the Institute for Microelectronics
at the Vienna University of Technology
- 2008 – 2012 Assignment as teaching assistant
- 2007 – 2010 Several research stays at Infineon Technologies AG, Munich, Germany
in the group of Hans Reisinger
- 2008 Visiting researcher at the Fudan University, Shanghai, China
in the group of Prof. Ming-Fu Li at the School of Microelectronics