Basel Halak *Editor*

Ageing of Integrated Circuits

Causes, Effects and Mitigation Techniques

Springer

Ageing of Integrated Circuits

Basel Halak
Editor

Ageing of Integrated Circuits

Causes, Effects and Mitigation Techniques

Springer

Editor
Basel Halak
The School of Electronics
and Computer Science
University of Southampton
Southampton, UK

ISBN 978-3-030-23780-6 ISBN 978-3-030-23781-3 (eBook)
https://doi.org/10.1007/978-3-030-23781-3

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To

my parents
as well as
Suzanne, Hanin, and Sophia
with love

Preface

The ageing of an organism in biology is defined as a progressive, irreversible process that inevitably ends with death. The maximal lifetime of an individual is significantly affected by ageing. The same is true for integrated circuits wherein ageing can be caused by several physical mechanisms, including bias temperature instability (BTI), hot carrier injection (HCI), and time-dependent dielectric breakdown (TDDB).

Ageing effects lead to a degradation in the performance and reliability of an electronic system, hence limiting its expected lifetime.

Variation-aware design techniques, such as conservative safety margins, can be used to reduce the impact of ageing on system reliability; however, the applications of such methods make it harder to develop competitive products and may lead to the elimination of performance gains of technology scaling. Therefore, there is a need for innovative approaches to improve the resilience of integrated circuit to ageing-induced failure without affecting its performance.

The prime objective of this book is to provide a timely and coherent account of the latest advances in the key research areas of IC's ageing; it has been developed as a collaborative effort among several international research groups, each providing an up-to-date summary of their latest findings and highlighting the remaining challenges and research opportunities. To facilitate the understanding of the material, each chapter includes a background section explaining related terminologies and principles, in addition to a comprehensive list of relevant references. The book is divided into three parts to enhance its readability, namely, physical mechanisms, mitigation techniques, and monitoring and adaptation approaches.

The Contents at Glance

This book explains the physical mechanism causing the ageing of integrated circuits, including a comprehensive analysis of its effects on the performance and reliability of integrated circuits. Afterwards, the book presents a number of mitigation techniques that can be applied at different stages of the life cycle of silicon chips. At the design stage, the book presents a synthesis algorithm that help produce ageing-resilient digital systems; in addition, it explores a number of application-dependent methods to improve system reliability. The book also discusses the state-of-the-art approaches for predicting ageing-induced failures and associated design adaptation techniques. More details on each chapter are provided below:

Part I: Ageing Physical Mechanisms and Effects

Chapter 1 provides a comprehensive review of the physical mechanisms causing the ageing of CMOS circuits.

Chapter 2 provides a detailed analysis of the impact of ageing on the reliability and performance of integrated circuits.

Part II: Ageing Mitigation Techniques

Chapter 3 presents an application-level solution to mitigate the impact of ageing on microprocessors using an anti-ageing software.

Chapter 4 discusses the impact of ageing on SRAM memories and review different approaches to mitigate against such effects.

Chapter 5 reviews the state-of-the-art techniques employed to enhance BTI lifetime reliability during digital synthesis.

Part III: Ageing Monitoring and Adaptation Techniques

Chapter 6 discusses the latest techniques used in ageing monitoring, including monitor designs and on-chip insertion methods.

Chapter 7 discusses the design of an ageing monitor to detect ageing-induced performance degradation in SRAM memories.

Chapter 8 presents a new design for a multipath delay monitor that used to predict ageing-induced timing errors.

Book Audience

The book is intended to provide a comprehensive coverage of the latest research advances in the key research areas of integrated circuits' ageing; this makes it a valuable resource for graduate students, researchers, and engineers working in these areas. I hope this book will complement the ongoing research and teaching activities in this field.

Southampton, UK
March 2019

Basel Halak

Acknowledgments

I would like to thank all of those who contributed to the emergence, creation, and correction of this book. First, I would like to thank my colleagues who have contributed chapters to this manuscript for taking the time to share their knowledge and for being very accommodating throughout the publication process. Special thanks go to the graduate students at Southampton, Grenoble, and Ioannina Universities for the many hours they have spent working in their labs to generate the experimental results. Of course, the book would not be successful without the contributions of many researches and expert in field of CMOS ageing.

Contents

Part I Ageing Physical Mechanisms and Effects

Part II Ageing Mitigation Techniques

Part III Ageing Monitoring and Adaptation Techniques

About the Editor

Basel Halak is the founder and director of the embedded system program at the University of Southampton and a fellow of the royal academy of engineering. He is currently a member of sustainable electronics and materials research group, as well as the cyber security group. He has written over 70 refereed conference and journal papers and authored two books, including the first textbook on physically unclonable functions. He received his PhD degree in Microelectronics System Design from Newcastle University and was then awarded a Knowledge Transfer Fellowship to develop secure and energy-efficient design for portable healthcare monitoring systems. He joined Southampton University in 2011 where he continued pursuing his research on developing reliable and secure systems. He has a long experience of the implementation flow of intergraded circuit from concept to silicon. His research expertise include the evaluation of security of hardware devices, the development of appropriate countermeasures and mathematical formalisms of reliability issues in CMOS circuits (e.g., cross talk, radiation, ageing), and the use of fault tolerance techniques to improve the robustness of electronics systems against such issues. He serves in several technical program committees such as HOST, IEEE IVSW, ICCCA, ICCCS, MTV, and EWME. Furthermore, he is an associate editor of *IEEE Access* and an editor of the *IET Circuits, Devices & Systems* journal and is also member of hardware security working group of the World Wide Web Consortium (W3C).

Part I
Ageing Physical Mechanisms and Effects

Chapter 1
Understanding Ageing Mechanisms

Domenik Helms

1.1 Introduction

Besides the pure functional correctness of an integrated circuit, IC designers at all times had to also regard other important design metrics, called *extra-functional properties*. The first two extra-functional properties, existing since the dawn of integrated circuits, were die area and circuit delay. The advent of power concerns in the 1980s introduced a third extra-functional property and made severe technology changes necessary such as the transition from NMOS to CMOS or the introduction of constant field scaling. In the 2000s, several new extra-functional properties, such as the various leakage currents, process variations, self-heating and IR drops, added a wide variety of new physical effects [19]. These new effects made several drastic architecture changes necessary, such as hafnium-based high-k oxides and multi-gate transistors. Even though being known for quite a long time [24, 41], ageing effects did not start to become relevant extra-functional properties before the 2010s. As their impact rises with each upcoming ever-smaller technology generation, there is so far no transistor architecture-based solution available.

At all times in the design of integrated circuits, an accurate prediction of the relevant extra-functional properties was the key for an effective design, keeping that property within its constraints. This is even more relevant for ageing effects: All former extra-functional properties, when poorly designed to be out of their constraints, failed in testing at the latest. Even though this might cause expensive design iterations, there is no in-field reduction in quality of service at stake. With ageing effects in contrast, even a careful testing cannot reduce the risk of an early system failure. Ageing aware design thus needs a reliable prediction of the

D. Helms (✉)
OFFIS – Institute for Information Technology, Oldenburg, Germany
e-mail: domenik.helms@offis.de

B. Halak, *Ageing of Integrated Circuits*,
https://doi.org/10.1007/978-3-030-23781-3_1

effects of ageing in order to optimize the system's performance at its end of life. Optimally, this ageing prediction is not just reliable (never underestimated) but also tight. Today's ageing models tend to vastly overestimate the ageing effects in many cases just to be on the safe side for some very rare occasions. This leads to huge overdesign efforts and thus finally to less performant or more expensive ICs.

Ageing-related and unrelated failures can be studied, when comparing the two most famous and most expensive hardware bugs, Intel shipped in their entire history: The most famous hardware bug ever occurred was surely the FDIV bug from 1994. Five incorrect table entries (thus an erroneous design) led to false division results. As this did only affect a small number of value pairs, this bug was not found, and the chip was shipped. The necessary callback did cost Intel some 475 M$. Even though Intel did drastic improvements in their test coverage afterwards, the Sandy Bridge bug in 2011, which was caused by an ageing effect, could affect some of the SATA driver transistors. The chips all passed testing and were shipped to the customers, failing then in field after a short time. This bug did cost Intel an estimated 700 M$.

1.2 Chapter Overview

The next section will introduce some basic concepts and give a brief overview on the observable ageing phenomenon, as well as over the fundamental working principles of MOSFET transistors. Section 1.4 will then detail the physics of oxide traps, explaining what oxide traps are, how they can be understood, how they can be activated and how they dominate the threshold voltage degradation of transistors. Section 1.5 then presents ageing models at all abstraction levels. Instead of explicitly modelling the trap physic, CET maps introduce a useful abstraction for virtually all ageing models. To enable modelling over years of operation time, models can additionally apply lifetime abstractions.

1.3 Basic Concepts

Before going into the details needed to understand and predict the ageing effects, this section will introduce the basic terminology and concepts for the later sections. Ageing will phenomenologically be introduced, and finally, the relevant working principles of MOSFET transistors will be discussed.

1.3.1 Terminology

At first, it is handy to introduce a clear definition of the terminology of the ageing-related effects, which might occur: First of all, an *error* in general is any kind of action, which is producing an incorrect result. This might include everything from a wrong decision made by a human designer, leading to a design error up to an automated productions step, not behaving exactly as anticipated, leading to a production error.

In contrast, a *fault*, which is also known as a *defect* is defined as an abnormal condition of a component. In the case of integrated circuits, faults are usually caused by a design or a production error, but an error does not necessarily have to lead to a fault. Faults changing the functional behaviour of the component are called *functional faults*, faults changing the extra-functional properties are called *parametric faults*. A typical example for a parametric fault is a too weak driving strength of a gate due to a design error or too high threshold voltage due to an error occurring in the oxidation process. Both errors will lead to an increase of the gate's propagation delay, which might either be within the tolerable margins or lead to an abnormal timing behaviour of this gate. An unintended electrical connection (short circuit) or a missing connection (open circuit) due to a designer's violation of design rules or due to process imperfections will lead to a change in the logic behaviour of a gate, thus a functional fault.

Finally, a *failure* is a deviation of the expected service of a component. Failures in general are caused by faults, but especially a parametric fault can also lead to no failure and then called a *masked fault*. In the above example, a gate timing fault may or may not lead to a timing deviation for the entire network, the gate is used in. Only if the gate lies on a path with critical or near critical timing, the overall network's timing is changed at all. Only if the timing is shifted outside the constraints, a timing failure will occur. Note that even though two individual timing faults do not have to cause a timing failure, both of them in combination might. It is a typical property of parametric faults that they do not lead to a direct failure where they occur but instead lead to a failure in a higher hierarchy: A too thick oxide is a parametric fault for a transistor, which leads to a parametric timing fault not a failure from the perspective of a single gate. For an entire netlist of gates building a *register transfer* (RT) component, strict timing bounds can be defined; thus the gate timing fault might or might not be translated into an RT timing failure.

In the context of integrated circuits, failures which are not caused by a design error are either coming from process variations (already existing at production) or from ageing (occurring over time). While *process variations* is a deviation between the planned design and a concretely produced chip due to production errors and prediction (thus design) errors, *ageing* is in the context of this chapter a deviation from the initial quality of a component due to the accumulation of faults over the lifetime of a component. As presented in Table 1.1, process variations and ageing can be further sub-divided:

Process variations can either be *random variations*, occurring in an unpredictable way per feature (transistor, metal line, isolation area); a *variation gradient*, effecting all features in a certain area in the same way, leading to hot spots, corner-to-corner deviations, die-to-die variations, wafer-to-wafer variations, lot-to-lot variations or even fab-to-fab variations; or *systematic variations*, occurring in the same way in each instance.

Table 1.1 Failure taxonomy and examples for each failure class

Failure				
Ageing		Process variations		
Degradation	Spontaneous	Random	Gradient	Systematic
Bias temperature instability	Electro migration	Random dopant fluctuation	Mask tilt	Sidewall ion scattering
Hot electron degradation	Time dependent dielectric breakdown	Line edge roughness	Thermal gradients at oxide formation	Optical proximity correction artefacts
...	...	...	...	...

An ageing failure caused by the occurrence of a single (usually functional) fault is referred to as *spontaneous failure*; an ageing failure caused by a drift in parameters due to the accumulation of parametric faults is called *degradation*. An ageing failure that, once occurred, will remain forever is called a *permanent failure*. If the failure can disappear, it is called a *transient failure* instead.

1.3.2 Ageing Phenomena

The most effective, earliest reported and thus most prominent ageing effect surely is the *negative bias temperature instability* effect (NBTI). NBTI occurs predominantly in PMOS (or p-type or p channel) transistors and causes an increase in the transistor's absolute threshold voltage. In analogue circuits, NBTI typically directly leads to a functional fault, which can accumulate to a functional failure over time. In digital circuits, it leads to a degradation of the switching speed, which does not directly lead to a failure but accumulates each time the system is under stress. Finally, NBTI can lead to a parametric transient timing failure for digital systems. Stress in the case of NBTI means that the PMOS transistor is in inversion; that means that its gate to body potential is substantially below 0 V for analogue circuits or at $V_{\mathrm{GB}} = -V_{\mathrm{DD}}$ for digital circuits. Higher voltages and higher temperatures both have an exponential impact onto the degradation, induced by NBTI. As soon as the stress is removed, the threshold degradation can regenerate, again with an exponential temperature dependence. Regeneration of NBTI obviously does not depend onto voltage levels, as the condition of stress removal is that V_{GB} is close to zero.

Especially for so-called high-k devices, also a similar effect, *positive bias temperature instability* (PBTI) can occur; predominantly in NMOS (or n-type or n channel) devices, PBTI has a similar degradation and healing behaviour and voltage and temperature dependency as NBTI. In fact, NBTI and PBTI are both manifestations of the same underlying process, even though the molecular basis for both effects is slightly different (see Sect. 1.4). High-k devices are typical for all technology nodes from 45 nm (introduced in 2007) and below. High-k devices replace most of the silicon oxide from the gate insulation layer by hafnium oxide for the sake of a high subthreshold slope (see Sect. 1.3.4) while keeping the gate tunnel effect [20] low.

Only PMOS NBTI and NMOS PBTI play an important role in terms of ageing. Nevertheless also the two further combinations of PMOS PBTI for high-k devices and NMOS NBTI can occur. As all four effects can be explained by the same mechanism just with changed polarity, in the remainder of this chapter, all four combinations are referred to as *bias temperature instability* (BTI).

The second important and also long-time known [41] effect is *hot carrier degradation* (HCI), which is often referred to as *hot carrier injection*. As finally it turned out that there is no injection taking place in this effect, we refer to it as hot carrier degradation but nevertheless use the much more common abbreviation HCI, trying to avoid confusion. Like BTI, HCI leads to an increase of the transistor's threshold voltage, affecting both NMOS and PMOS devices. HCI seems to have just a weak temperature dependency. Instead, HCI dominantly occurs under a certain combination of drain and gate voltages [5]. In fact the difference in thermal and especially voltage dependency is the best option to separate both effects in measurement [38]. Unlike BTI, it seems to be no or just little recovery for this effect [43]; thus HCI and BTI both contribute to the same parametric permanent timing failures for digital systems and functional failures for analogue systems. It is assumed, that BTI and HCI can both be explained by the same molecular faults but with different electrical activation mechanisms.

Even though *random telegraph noise* (RTN) is rather a noise than an ageing effect, it is nevertheless closely linked, especially to the BTI effect. A substantial part of the RTN effect can be explained by the BTI effect degrading and regenerating on very short time scales [48]. BTI is caused by discrete events (see Sect. 1.4), each with their individual typical timing for degradation and regeneration [15]. The shortest of these events are distributed over the μs range. Their frequent degradation and regeneration can explain the RTN behaviour. In that sense, RTN is similar to BTI, just occurring on shorter time scales.

In *time-dependent dielectric breakdown* (TDDB), disturbances (conductive faults) in the transistors oxide silently sum up over a long period of operation time. At first the observable tunnelling current rises in discrete steps, which is called a soft breakdown, causing parametric power and timing faults as the gate tunnelling leakage rises and thus the charging of a fan-in slows down. Finally, the gate tunnelling current enters a runaway situation in which the drastic increase in tunnelling current results in a short circuit of the gate oxide and thus a spontaneous permanent failure of the gate [6].

In *electro migration* (EM), high currents and high temperatures in the interconnect structures can lead to a relocation of the interconnect metal, again leading to open circuits (where metal is lost) and short circuits (where the metal is accumulated) [27]. Even though these effects are caused by an accumulation of faults, they lead to almost no noticeable parameter degradation until the component spontaneously fails, thus leading to spontaneous permanent failures.

Ionizing radiation, striking the transistor can ignite the latch-up effect, causing two parasitic bipolar transistors inside the MOSFET structure to amplify each other leading to uncontrollable supply to ground current flow and finally the thermal destruction of that gate, thus clearly leading to spontaneous permanent failures. Ionizing radiation might also cause a glitch, a transient change of a transistor's output voltage. In sequential circuits such as latches, registers and memory cells, these glitches can also cause the sequential loop to switch its state. This effect can cause permanent failures from a data point of view but only leaves a transient fault from a hardware perspective.

Besides these microscopic ageing effects, occurring in the fundamental structures of an embedded system, there are also many macroscopic effects such as thermal cracking, occurring due to diverging thermal expansion of different materials or delamination. These macroscopic material science effects are not in the scope of this chapter.

1.3.3 Mission Scenarios

As described in Sect. 1.3.1, while producing integrated circuits, production defects, imperfections and process variations may occur, which may either directly lead to a failure (then accounted for as yield loss) or may never lead to a failure (then called masked errors) or may lead to a spontaneous failure sometime within the lifetime of the component. Over-proportionally many of these spontaneous failures are activated very soon after the first usage of the component. Due to various effects, integrated circuits suffer from a high infant mortality. Thus, the systems are usually applied to high stress conditions for a short time (called burn-in) before testing in order to be able to provoke these effects and screen for early deaths.

Over the entire lifetime of the system, the system may fail with a low but finite probability due to the spontaneous failures such as latch-ups. After a certain mission-dependent time, the probability for a parametric failure caused by a degradation mechanism will start to exceed the failure rates for the spontaneous failures. In total, this results in a characteristic bathtub distribution (Fig. 1.1).

The point in time, where this *wear-out* begins, and generally the entire degradation behaviour, both are determined by two factors: Firstly, the initial condition of the system will mainly due to production errors (process variations) result in an individual distribution of faults, directly or indirectly influencing the degradation behaviour per transistor. For instance, the distribution and structure of defects in the oxide are directly influencing the BTI timing; the distribution of dopants in the channel then influences the impact of each oxide defect onto the device's threshold voltage. Secondly, ageing strictly depends on the dynamically changing operation conditions, especially voltage levels, temperatures and duty factors. A duty factor describes the ratio of time in which a digital system is under stress. Due to power

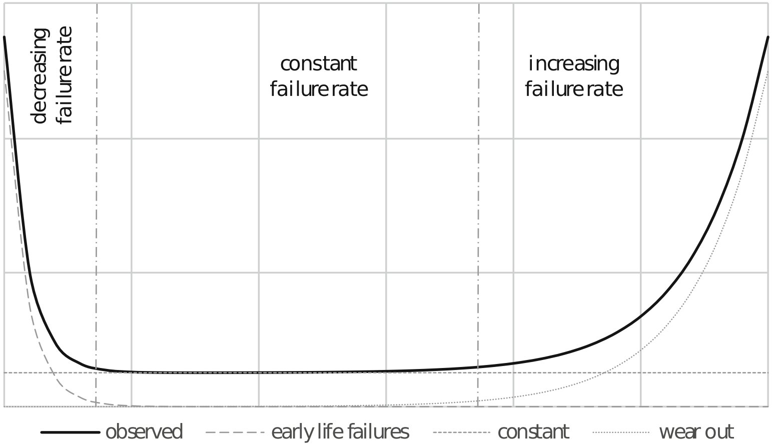

Fig. 1.1 The three major contributors to the integrated circuit failure probability over time are the early life failures, quickly reducing in probability over the first hours or days of operation; failures with a constant failure probability, such as radiation induced failures; and wear-out failures due to ageing

management techniques such as dynamic voltage and frequency scaling or power gating, due to typical idle phases and due to effects such as signal correlation, self-heating and IR drop, these quantities (temperature, voltage and duty factor) are neither constant nor unpredictable. In contrast, they have a vast impact onto the pace of degradation.

In order to assess the ageing behaviour of an integrated circuit, either by measurement, by simulation or by modelling, these factors have to be understood and well defined. If the expected lifetime of the system, as well as the history of temperature, voltage and duty factor for the system, is specified, this specification is referred to as a *mission scenario*. In the simplest case, the mission scenario just states worst-case assumptions, asking to measure, simulate or model the ageing behaviour of a system, which is for its entire lifetime at the worst-case end of its specifications (i.e. highest temperature, highest voltage, always under stress).

For some applications, with highest reliability demand, this assumption might be a good choice even though it obviously leads to a drastic over-prediction of the ageing and thus to reduced system quality and/or increased design cost. At and below moderate reliability requirements, it is possible to loosen the worst-case assumptions to some degree in order to increase quality or reduce cost. For instance, a car manufacturer might shoot for 20 years of failure-free regular operation but only if the system shows a typical pattern of usage, such as being active for less than 2 h per day on average. For more heavy usage, their lifetime requirements might be lower. When specifying these requirements into the mission scenario such as worst-case temperature of 175 °C (from ambient, motor heat and circuit self-heating) only for 2 h a day and worst-case ambient temperature of 55 °C for the other 22 h as well as a duty factor of 2 h a day, a much tighter ageing assessment can be made, leading to less overdesign, less design costs and/or higher quality.

1.3.4 MOSFET Transistor Basics

This section aims at giving the reader the background in transistor physics, needed to understand the context of the following sections, especially Sect. 1.4. Even though there are various transistor technologies available, virtually all transistors used today are *metal oxide semiconductor field effect transistors* (MOSFET). The basic working principle of a MOSFET transistor is that a silicon lattice is doped to become a majority carrier conductor at both ends (source and drain) and a minority carrier conductor in between (channel). Majority carrier means an electron in case of an NMOS transistor and a hole in case of a PMOS transistor; minority carrier refers to the opposite polarity, respectively. The gate is above to the channel, only separated by an isolating layer, called the oxide layer. In the same way that the doping can control the ratio of free electrons and holes (and thus which of them becomes the majority carrier), also an electric field, applied by a gate voltage, can change this ratio. The working principle of MOSFET now is that the gate field acts on the channel in exactly the opposite way, the doping did, thus converting the minority carrier dominated channel into a majority carrier conductor. This process is called *inversion*, and it exponentially increases the amount of majority carriers, available for the source to drain current.

The rate at which the inversion takes place is described by the source-drain current equation [35]:

$$I_{\mathrm{SD}} = I_0 \cdot e^{\frac{V_{\mathrm{GS}} - V_{\mathrm{th}}}{n V_{\mathrm{T}}}} \tag{1.1}$$

where n is called subthreshold slope which can be computed as

$$n = 1 + \frac{C_{\mathrm{dep}}}{C_{\mathrm{ox}}}, \tag{1.2}$$

where C_{ox} is the capacitance between gate and channel and C_{dep} is the capacitance between substrate and channel. The subthreshold slope is the fundamental problem of today's MOSFET transistors in general: In order to switch on and off quickly without huge power consumption, I_{SD} needs to be reduced and increased as much as possible by as little V_{GS} voltage swing as possible. Therefore the subthreshold swing n needs to be as small as possible. From Eq. 1.2 we can immediately see that we therefore need to make the channel to substrate capacitance as small and the oxide capacitance as large as possible. We can also see from Eq. 1.2 that n can never become less than one. That means that at room temperature

$$T = 300\,\mathrm{K} \Rightarrow V_{\mathrm{T}} = \frac{k_{\mathrm{B}} T}{e^{-}} = 1.38 \cdot 10^{-23}\,\mathrm{J/K} \cdot 300\,\mathrm{K}/1.60 \cdot 10^{-19} As \approx 25.85\,\mathrm{mV} \tag{1.3}$$

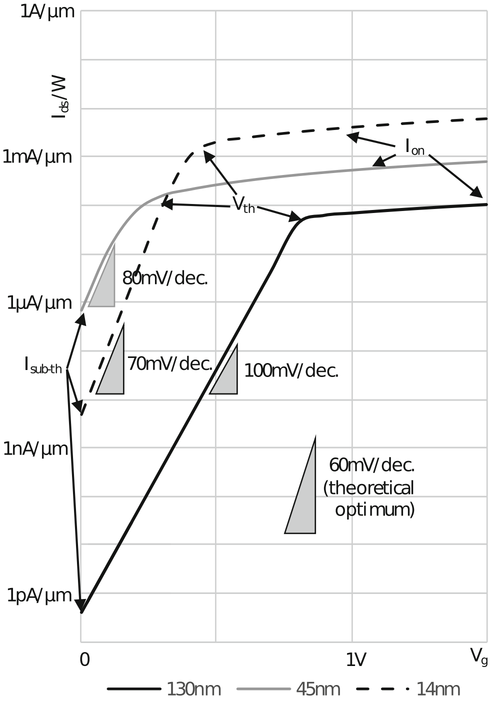

Fig. 1.2 Qualitative behaviour of a 130 nm SiO_2 bulk, a 45 nm high-k, and a 14 nm FinFET technology: The higher gate capacitance from the HfO_2 oxide already resulted in a substantial decrease in subthreshold slope. The multi-gate structure of the FinFET also reduces C_{dep} while further increasing C_{ox}. The coming nanowire devices might even approach the theoretical slope limit of 60 mV/dec. Together with the threshold voltage, these two parameters dominate the transistors operation speed (I_{on}), standby power (I_{off}) and efficiency (I_{on}/I_{off})

Thus, without knowing any other transistor parameter and assuming optimal values for C_{ox} and C_{dep}, we can deduce that an increase of 25.85 mV in V_{GS} is at least needed to rise the drain to source current by a factor e or equivalently V_{GS} and must rise by a factor

$$25.85\,\text{mV} \cdot \ln(10) \approx 60\,\text{mV} \tag{1.4}$$

to let I_{DS} rise by a factor of ten (see Fig. 1.2).

This slope limitation ultimately caused the end of voltage scaling back in 2003 as the threshold voltage should not be below 3–4 times the subthreshold slope to ensure a factor 1000–10,000 I_{on} to I_{off} ratio. The slope limitation is also the main motivation behind all major technology changes such as high-k devices (increasing C_{ox} while keeping tunnelling currents under control) and multi-gate devices (increasing C_{ox} and decreasing C_{dep} at the same time). In the near future, this slope limitation may even cause a technology change away from MOSFET towards so-called steep slope devices. Most promising candidates here are tunnel FETs, being theoretically able to reduce the subthreshold swing to below 1 mV and being demonstrated to actually do below 32 mV[40], allowing stable operation at 0.1 V supply voltage.

For our recent 3D multi-gate MOSFET devices, the unavoidable process variations are already spoiling deeper scaling, as variations in the cutting edge threshold voltage lead to extreme power and timing issues. In these deeply scaled devices, the charge of a single electron or hole, when trapped in the oxide, will lead to a measurable change of the effective threshold voltage of the entire device. This is exactly what causes ageing: charges can get trapped in the oxide and will then accumulate over time. For the most dominant effects, this leads to an increase in threshold, most dominantly for the PMOS devices, which then leads to lower I_{on}, which again leads to longer transition times.

To conclude this section: As MOSFET devices have a finite subthreshold slope, their power and timing behaviour is extremely vulnerable towards variations in the threshold voltage. Ageing occurs due to traps in the oxide, modifying the threshold voltage. In the following chapter, we will thus study oxide traps in more detail and will understand that knowing the state of each trap in each transistor at any time completely specifies the ageing behaviour of our system.

1.4 Oxide Defects

Despite the large variety of ageing effects introduced in Sect. 1.3.2, it turns out that most of the relevant effects, such as NBTI, PBTI, HCI, RTN, and TDDB, all can be explained by imperfections in the gate oxide. Thus this section discusses the recent understanding of formation and activation of these oxide traps in detail.

1.4.1 Trap Formation

The first observations ever made on any ageing effect were already reported in 1977. There is a systematic upshift in the threshold voltage of PMOS devices over their lifetime [24]. This effect, at first referred to as *negative bias stress*, was since then observed in all PMOS devices. In those early days, the technology structures were large, and the BTI behaviour was very smooth as presented in Fig. 1.3. From this figure, it is obvious that the BTI effect depends on both the temperature and the voltage and it leads to an increase in threshold voltage under stress and a decrease once the stress is removed.

The first explanation at hand for this phenomenon was the *reaction-diffusion* model (RD), which was already proposed by [24]. The RD model is explained by the hydrogen atoms passivating the channel oxide barrier [31]. Hydrogen naturally occurs at the oxide to channel interface. It is used to passivate dangling bonds stemming from the slight lattice size mismatch between the channel's silicon and oxide's silicon oxide. If a dangling bond is not passivated, it leads to a positive charge in the interface, which will effectively counteract the field from the gate – thus increasing the threshold voltage. The RD model now proposes that positive charges can enter the silicon-hydrogen bond under stress, leaving a silicon with only three bonds to other silicon atoms and a positive charge due the one missing

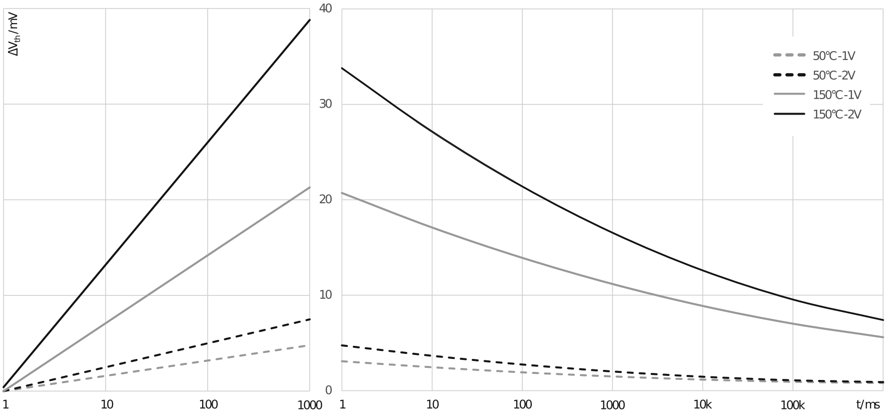

Fig. 1.3 Influence of temperature and voltage onto threshold degradation and regeneration

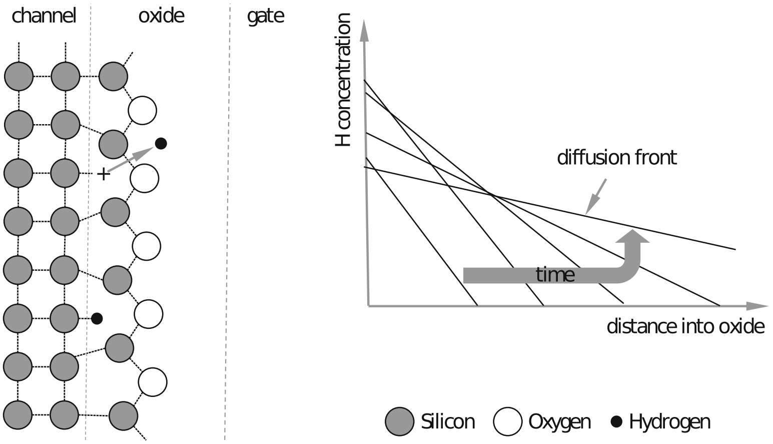

Fig. 1.4 Reaction-diffusion model for BTI. Trapped holes can weaken the bonds of the hydrogen passivation in the channel interface. Over time, the diffusion front moves under stress

electron, which used to be shared with the hydrogen to form this bond. The now free hydrogen can get lost, reacting with another hydrogen ion to form a neutral hydrogen molecule. The neutral hydrogen can then diffuse away (see Fig. 1.4), as it is no longer affected by any fields, holding it back. In RD, healing was explained by hydrogen, recombining with the defects if the stress is removed soon enough before the hydrogen diffused away for good. Once the hydrogen diffused away a certain distance, the open bond remains open and leaves a permanent damage.

Even though the predictions of the RD model suited the observations of the early technologies well, measurements using more recent transistors showed significant deviations. RD completely failed to explain the time constants, observed for small devices. In 2009, the alternative *trap charge* model (TC) was thus proposed [14].

In the TC model, BTI is explained by molecular defects inside the oxide. In a first version, it was assumed that oxide vacancies (defects, where silicon directly bonds to another silicon without an oxygen bridge) cause such defects. The initial model also proposed a two-state process [14], where the traps in the oxide are initially uncharged (state 1) and can capture and store a positive charge (state 2). Even though this model could much better explain the measured time constants, it still stood in contradiction with the measured distributions of capture and emission times. In [15], the model was extended to a four-state model (equivalent to Fig. 1.5a, b, d, e) but with oxygen vacancies instead of hydroxyl groups as follows: The neutral state (Fig. 1.5a) can trap a charge under stress, which will usually discharge again after a short time. In the trapped state (Fig. 1.5d), a silicon-oxygen bond is weakened,

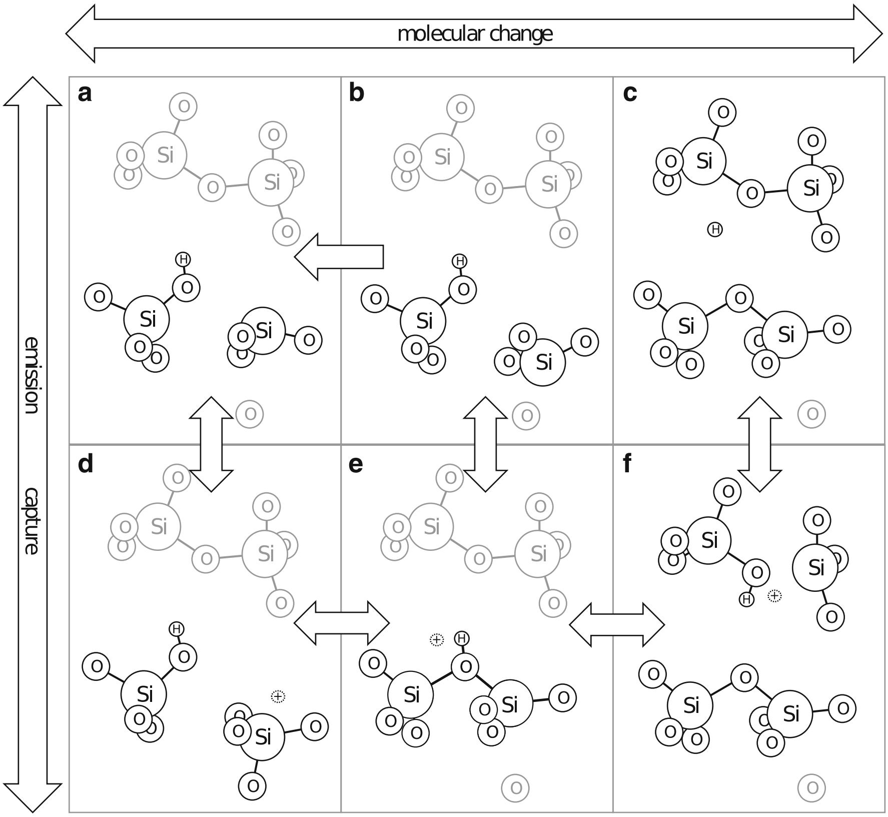

Fig. 1.5 Charge traps in the oxide are formed by lattice imperfections. In the initial configuration (**a**), a single hydrogen atom might cause a deviation from the conventional silicon-oxygen-silicon structure. Under stress, holes will be frequently captured and reemitted (a⇌d). Elevated temperatures increase the probability for a molecular change of the oxide (d⇌e). Depending on the exact molecular arrangement, (**e**) might be stable over long periods of time. Transitions such as e⇌f⇌c can also explain, why certain traps seem to disappear and reappear over larger time spans

facilitating a structural (molecular) transformation into another quasi-permanent state (Fig. 1.5e). In state (e), a positive charge is trapped inside the oxide, leading to an increase of the threshold voltage in PMOS. Once the stress is removed, a trapped charge (Fig. 1.5e) can then either undergo a structural relaxation (Fig. 1.5d) first and discharge to the neutral state afterwards (Fig. 1.5e) or discharge first (Fig. 1.5b) before the structural relaxation. This four state model could already explain oxide traps with individual capture and emission times.

When measuring BTI in silicon, there is another effect which is in contradiction to this TC model [15]: Individual traps in the oxide can be observed over a period of time, after which they seem to disappear. After some time, up to weeks of absence, the traps then reappeared with exactly the initial behaviour. In [48], the TC model was thus extended again, resulting in a six-state model, as displayed in Fig. 1.5. Other molecular disturbances than oxide vacancies, such as hydroxyl groups or hydrogen bridges, are analysed in [48]. They then added two additional states the system can pass, finally ending in a neutral quasi-permanent state (Fig. 1.5c), explaining the disappearance of a trap.

As there was strong evidence that hydrogen plays an important role in BTI, the RD model was favoured by many research groups, even after the introduction of the TC model and even though the time constants observed in BTI could better be explained by the TC model. Detailed measurements at single traps showed time constants up to weeks, while small devices with just 1–2 nm thick oxides can explain time constants in the order of milliseconds [15]. As of today, this dispute is settled [16], with the updated version of the TC model, where traps are formed not by oxide vacancies but by hydrogen bridges (silicon-hydrogen-silicon bonds) or hydroxyl groups (silicon-oxygen-hydrogen-silicon bonds) inside the silicon oxide [48]. As this new model fits well to all observed NBTI and PBTI data for NMOS and PMOS, most scientists today agree on the TC model. Nevertheless, there are some recent publications, still referring to RD.

1.4.2 Trap Activation

Section 1.4.1 explained how charges can be captured and released inside the gate oxide and how this qualitatively influences the threshold voltage of large devices in a seemingly continuous way. Today's transistors are very small, the density of these oxide charge traps is not too high in silicon oxide, and they are randomly distributed. Thus, when screening small enough transistors on a test chip in a decent technology, it is possible to find single transistors having just a few or even just one single trap in their oxide. When finding such a transistor, it is possible to study the behaviour of each individual trap in great detail. Such an analysis is then called time-dependent trap spectroscopy [15].

As sketched in Fig. 1.6, it seems as if each single trap has a fixed contribution to the threshold voltage. When analysing devices of different sizes, the absolute shift (in mV) of the threshold voltage tends to become larger with smaller device