ACSP · Analog Circuits and Signal Processing

Elie Maricau
Georges Gielen

Analog IC Reliability in Nanometer CMOS

Springer

Analog Circuits and Signal Processing

For further volumes:
http://www.springer.com/series/7381

Elie Maricau · Georges Gielen

Analog IC Reliability in Nanometer CMOS

Springer

Elie Maricau
ESAT-MICAS
KU Leuven
Heverlee
Belgium

Georges Gielen
ESAT-MICAS
KU Leuven
Heverlee
Belgium

ISBN 978-1-4614-6162-3 ISBN 978-1-4614-6163-0 (eBook)
DOI 10.1007/978-1-4614-6163-0
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012954854

Printed on acid-free paper

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

Abstract

Today, micro-electronic circuits are undeniably and ubiquitously present in our society. Transportation vehicles such as cars, trains, buses, and airplanes make abundant use of electronic circuits to reduce energy consumption and emission of greenhouse gases and to increase passenger safety and travel comfort. Other products using electronic circuits are smartphones, tablet PCs, game consoles, household appliances, satellites, base stations, servers, etc. Each of these applications is becoming increasingly more complex to build. At the same time, the quality and reliability requirements for electronic circuits are more demanding than ever.

To guarantee a high production yield and a sufficient circuit lifetime, possible hazards and failure effects have to be considered throughout the entire design flow. Such a flow includes the initial concept, the design itself, the testing of the prototype circuit, and finally the production process. The majority of integrated circuits manufactured today is processed in a complementary metal-oxide semiconductor (CMOS) technology. To reduce cost and to increase performance, the dimensions of all circuit components are shrinked with each new technology node. Associated with this technology scaling are the atomistic size of modern transistors, an increase of the gate-oxide electric field, and the introduction of new gate and channel materials. The combination of these elements results in an emerging reliability problem for advanced nanometer CMOS technologies. Transistor wearout manifests itself as a gradual and time-dependent shift of circuit characteristics which can result in circuit failure. Especially analog circuits, which are typically used as an interface between the real world and a digital backend, can be very sensitive to such small circuit parameter variations.

This work focuses on the simulation and analysis of analog circuit reliability. The models and simulation techniques proposed in this dissertation are aimed to serve as an aid for circuit designers to better understand the impact of aging effects on their circuits and to enable the development of failure-resilient design solutions.

In a first part of the work, an overview of all relevant nanometer CMOS unreliability effects is given and transistor compact models for the most important

aging effects are proposed. A distinction between spatial unreliability effects, resulting from process variations, and temperoral unreliability effects, which are time-dependent, can be made. The latter can again be divided into transient effects such as noise and electromagnetic interference, and aging effects such as breakdown, bias temperature instability, and hot carrier injection. This work primarily concentrates on the aging effects. To enable efficient and accurate circuit lifetime simulations, transistor compact models for each aging effect are proposed. These models include the most important circuit-related stress parameters such as voltages, transistor dimensions, and temperature. Important effects such as partial recovery of the transistor damage when the stress voltage is reduced, are also supported. Each model is validated with measurements. Also, models for stochastic aging effects in sub-45 nm CMOS, which result in time-dependent transistor mismatch, are discussed.

A second part of the book focuses on the development of efficient simulation methods to analyze the impact of transistor aging on an entire circuit. Existing reliability simulators, published in the literature or commercially available, still suffer from a lot of deficiencies. Often, these tools do not support all unreliability effects and especially the impact of process variations and stochastic aging effects is in most cases not included. The tool set presented in this work aims to solve these problems, while still limiting the computational effort. The proposed simulator includes support for all important deterministic and stochastic aging effects. Further, the interaction between process variations and aging effects can be analyzed and visualized. In addition to a visualization of the time-dependent performance shift of the circuit under test, reliability weak spots can be detected. This enables a designer to search for dedicated solutions in case of a reliability problem. To limit the simulation time, the simulator uses a response surface method which models the time-dependent circuit performance based on only a limited set of SPICE-based reliability simulations. Finally, a hierarchical simulation framework based on an adaptive sample selection algorithm and a nonlinear symbolic regression algorithm enables the reliability simulation of large analog circuits within a reasonable time frame. Each part of the simulation framework is demonstrated on an example circuit.

The last part of this work applies the proposed reliability compact models and simulation methods to a set of commonly used analog circuits. Factors that determine the circuit lifetime are explored and illustrated with examples. Further, a design for reliability flow is demonstrated on an example IDAC circuit resulting in the design of a reliable circuit with minimum guardbanding. Finally, the lifetime of small- to medium-sized digital circuits is investigated. Although, the methods proposed in this work are primarily intended for analog circuits, they are also applicable to small- and medium-sized digital circuits when these are defined as a SPICE netlist.

The models and simulation techniques developed in this work are intended as a first step towards understanding the impact of transistor aging on analog integrated circuits. Eventually, this understanding can help designers in designing guaranteed reliable and robust circuits in future CMOS process nodes.

Contents

Abbreviations

AC	Alternating Current
ADC	Analog-to-Digital Converter
AEC	Automotive Electronics Council
ALT	Accelerated Life Test
AMS	Analog and Mixed Signal
BERT	Berkeley Reliability Tools
BSIM	Berkeley Short-channel IGFET Model
BTI	Bias Temperature Instability
CDF	Cumulative Density Function
CMOS	Complementary Metal-Oxide-Semiconductor
CHE	Channel Hot Electron
CP	Charge Pumping
DAC	Digital-to-Analog Converter
DAHC	Drain Avalange Hot Carrier
DC	Direct Current
DFR	Design For Reliability
DPT	Design Patterning Technology
DUT	Device Under Test
eMSM	extended Measurement Stress Measurement
EM	ElectroMigration
EMI	ElectroMagnetic Interference
ENOB	Effective Number Of Bits
EOT	Effective Oxide Thickness
EUVL	Extreme Ultraviolet Lithography
ESD	ElectroStatic Discharge
FET	Field Effect Transistor
FF	Fractional Factorial
FFX	Fast Function eXtraction
FMEA	Failure Mode Effect Analysis
FTC	Fast Transient Charging
GBW	Gain-BandWidth

HALT	Highly Accelerated Life Test
HASS	Highly Accelerated Stress Test
HBD	Hard Breakdown
HCI	Hot Carrier Injection
HKMG	High-K Metal Gate
IC	Integrated Circuit
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
IL	Interfacial Layer
ITRS	International Technology Roadmap for Semiconductors
LEM	Lucky Electron Model
LER	Line Edge Roughness
LHS	Latin Hypercube Sampling
LWR	Line Width Roughness
MARS	Multivariate Adaptive Regression Splines
MC	Monte Carlo
MOSFET	Metal Oxide Semiconductor Field Effect Transistor
MT	Multiple Trapping
MTTF	Mean Time To Failure
NBTI	Negative Bias Temperature Instability
NF	Noise Figure
nMOS	n-type MOS Transistor
NMSE	Normalized Mean Square Error
OPC	Optical Proximity Correction
OTF	On The Fly
PBD	Progressive Breakdown
PBTI	Positive Bias Temperature Instability
PCB	Printed Circuit Board
PDF	Probability Density Function
pMOS	p-type MOS Transistor
PVT	Process Voltage Temperature
RD	Reaction Diffusion
RDD	Reaction Dispersive Diffusion
RDF	Random Dopand Fluctuation
RF	Radio Frequent
RMS	Root Mean Square
RTL	Register Transfer Level
RTN	Random Telegraph Noise
SBD	Soft Breakdown
SEU	Single Event Upset
SGHE	Secondary Generated Hot Electron
SHE	Substrate Hot Electron
SNR	Signal-to-Noise Ratio
SPICE	Simulation Program with Integrated Circuit Emphasis
SR	Symbolic Regression

SSN	Simultaneous Switching Noise
SSPA	Switching Sequence Post Adjustment
SVM	Support Vector Machines
TDDB	Time Dependent Dielectric Breakdown
TFR	Test For Reliability
TTF	Time To Failure
UDRM	User Defined Reliability Model
VCO	Voltage Controlled Oscillator

Symbols and Quantities

A_{VTH}	Mismatch technology parameter
C_{ox}	Oxide capacitance
E_a	Activation energy
E_{lat}	Lateral electric field
E_{ox}	Oxide electric field
g_{o}	Small signal output conductance
I_{DS}	Drain-source current
I_{GS}	Gate-source current
I_{sub}	Substrate current
k	Boltzmann constant (1.3806503×10^{23} J/K)
k_{F}	BTI forward dissociation rate
k_{R}	BTI annealing rate
L	Transistor length
N_{IT}	Number of interface traps
N_{OT}	Number of oxide traps
q	Electron charge (1.602×10^{-19} C)
t_{BD}	(Mean) time to breakdown
t_{ox}	Oxide thickness
T	Temperature
T_{str}	Total stress time
V_{BS}	Bulk-source voltage
V_{DS}	Drain-source voltage
V_{GS}	Gate-source voltage
V_{str}	Stress voltage
V_{TH}	Threshold voltage
W	Transistor with
β	Carrier mobility
Δ	(Aging induced) parameter shift
δ	(Process induced) parameter variation
$\varepsilon_{\mathrm{ox}}$	Oxide permittivity
$\mathcal{F}$	Factor space

$\phi_{\mathrm{IT,e}}$	Critical energy for electrons to create an interface trap
λ_e	Hot-electron mean free path
μ	Mean
$\mathcal{P}$	Performance space
σ	Standard deviation
ϑ	Relation between factor space and performance space

Chapter 1
Introduction

1.1 Introduction

This work aims to provide the reader with a comprehensive understanding on the subject of modeling, analyzing and understanding the impact of transistor aging on analog integrated circuits (IC) in a nanometer complementary metal-oxide-semiconductor (CMOS) technology. The first chapter of this work introduces the problem studied and the major subjects addressed in this book.

The chapter is outlined as follows. First, a brief history of reliability engineering is given in Sect. 1.2. Then, Sect. 1.3 discusses the importance of reliability in electronic systems today. Depending on the application, different circuit requirements need to be fullfilled, but reliability is always a key element. The focus of this work is on integrated circuits processed in an advanced nm CMOS process. The evolution of the CMOS production process, from the conventional SiO_2 process to high-k metal-gate (HKMG) devices used in sub-45 nm technologies and the introduction of new device architectures such as FinFETs for sub-28 nm technologies, results in an increasing amount of reliability problems and is discussed in Sect. 1.4. Section 1.5 then discusses the basic activities needed to guarantee the reliable operation of a product or system. Also, the currently established IC design for reliability strategy is reviewed. Finally Sect. 1.6 presents the chapter conclusions.

1.2 Reliability Engineering: A Brief History

Reliability is a popular concept and is also seen as a commendable attribute of a person or an object. Etymologically, the term stems from the Scottish word *raliabill* (rely+able, sixteenth century) (Saleh and Marais 2006). The first recorded usage of the word reliability, albeit referring to a person instead of an object, dates back to 1816 when it was introduced by poet Samuel Taylor Coleridge (Saleh and Marais 2006). In praise of his friend Robert Southey, Coleridge wrote:

E. Maricau and G. Gielen, *Analog IC Reliability in Nanometer CMOS*,
Analog Circuits and Signal Processing, DOI: 10.1007/978-1-4614-6163-0_1,

> *He inflicts none of those small pains and discomforts which irregular men scatter about them and which in the aggregate so often become formidable obstacles both to happiness and utility; while on the contrary he bestows all the pleasures, and inspires all that ease of mind on those around him or connected with him, with perfect consistency, and (if such a word might be framed) absolute reliability.*

Today, the term is used extensively by the general public and the technical community. Web of science, for example, lists over 260,000 technical papers with 'reliability' as a keyword, while the popular search engine Google even returns over 218,000,000 hits (Saleh and Marais 2006). When referring to an object, reliability is now defined as the ability of a system or component to perform its required functions under stated conditions for a specified period of time. Further, between 1816 and now, many industrial evolutions such as the development of the first airplane in 1903 and the introduction of the Ford model T in 1908 as first mass produced car took place. All these inventions and technical ideas contributed to the rise of reliability engineering as a scientific discipline in the early 1950s (Fig. 1.1).

Probability theory and statistics are the essential ingredients without which reliability engineering could not have emerged. The theory of probability was established in 1654 by Blaise Pascal and Pierre de Fermat in the context of gaming and gambling. In 1812 Laplace introduced a series of techniques, related to probability and statistics, expanding their scope to various other problems such as demographics, population estimation and life insurance. Another essential enabler for the rise of reliability engineering as a technical discipline was the idea and practice of mass production as a means for cost reduction. To deal with quality issues in high-volume production,

Fig. 1.1 Radar will win the war (Life Magazine 1944). Vacuum tubes, used as active components in World War II radar systems during the so-called wizard war, were the major source of system failure and initiated reliability engineering efforts in the early 1950s

first came statistical quality control in the late 1920s. Later, reliability engineering was introduced in the mid 1950s, to deal with the (un)reliability of the vacuum tube. The tube, which was invented by Lee de Forest in 1906, initiated the electronic revolution enabling a series of applications such as the radio, the television and the radar. In World War II, electronics played a critical role and contributed to the allies winning the 'wizard war'. The vacuum tube, an active element that was part of the radar systems, was however also the major source of equipment failure. This prompted the US Department of Defense (DoD) to initiate a series of studies to look into these failures after the war. In 1952, the advisory group on reliability of electronic equipment (AGREE) was jointly established between the DoD and the American electronics industry. Its mission was to recommend measures that would result in more reliable equipment, to help implement reliability programs in government and civilian agencies and to disseminate a better education on reliability. The first conference on quality control and reliability of electronics was held in 1954 and its proceedings evolved into a journal that is still being published: IEEE Transactions on Reliability. After the consolidation of the initial efforts addressing reliability issues in components, various branches with increased specialization were founded throughout the 1960s and 1970s. These branches include the development of improved statistical techniques (redundancy modeling, Bayesian statistics, etc.), modeling of physical causes of failure, prediction of reliability at a system level, etc.

The chain of events described above eventually led to the practical reliability engineering as it is known today. In the next sections, the focus is on the reliability requirements set by modern electronic systems and the problems involved with guaranteeing reliability in nanometer CMOS processes.

1.3 Reliability of Electronic Systems

Many applications in our society today make use of advanced micro-electronic circuits. Depending on the application, each of these circuits has different requirements (see Fig. 1.2):

1. Consumer products such as cellphones, TVs and computers represent a huge market with a lot of competition. Here, a short time-to-market and a first-time-right approach are required to maintain or increase market share.
2. Safety-critical applications require reliable circuit operation with a lifetime of ten to twenty years. Examples of such applications are a sensor interface used to monitor vital parameters in a car or biomedical products such as a pacemaker.
3. A third type of integrated circuits is used in very harsh environments. Circuits embedded in automotive products or airplanes, for example, typically have to endure large temperature variations and electromagnetic interference. Further, ICs used in more extreme applications such as satellites or sensors in nuclear reactors suffer from radiation effects. Each of these circuits, however, always needs to operate correctly.

Fig. 1.2 Micro-electronic circuits are used in many applications. Each of these applications has specific needs. For consumer electronics the time to market is essential. Safety-critical products require very high-quality components. Circuits used in harsh environment need to operate in extreme conditions

Note how a circuit can belong to multiple categories at the same time. ICs for automotive products, for example, typically serve a large and competitive consumer market and are used in safety-critical and harsh environments at the same time. Designing these ICs is therefore very challenging.

Further, this vast market has lead to products with increasing functionality at a lower cost. To reduce production costs, the semiconductor industry continues to scale transistor devices to smaller nm CMOS technologies. To maintain the effective performance scaling, however, the oxide electric fields and current densities are increasing continuously. Advanced CMOS nodes are now reaching values where it is very challenging to guarantee–by the mere production and limiting the usage to predefined boundaries (e.g. maximum supply voltage)–reliable circuit operation over the intended product lifetime. In industry, reliability issues can therefore result in a stall of the production process or even a recall of already sold products. In 2011, for example, Intel was about to launch their Sandy Bridge processor when a potential reliability problem was detected (Clarke 2011). The problem, which was not spotted during extensive functional testing, was a gradual performance reduction and even total failure of the serial-ATA channels in about 5 % of the manufactured ICs. Another, unfortunately more common, problem with wireless electronic products is battery reliability. In 2009, HP recalled a large number of laptops after a number of incidents

with the battery that overheated and ruptured (Ogg 2009) due to a malfunctioning electronic protection circuit. A similar problem with overheating of Sony laptops was reported in 2010 (Lipka 2010).

The example cases above demonstrate that it is very difficult, even for a major company, to guarantee absolute reliability over a product lifetime. Nevertheless, even a limited number of product failures can already result in excessive warranty costs and severe brand damage. Further, electronic systems contain mixed-signal circuits with embedded high-performance analog or mixed-signal blocks and potential sensitive RF frontends. Although digital building blocks also suffer from aging effects, they are inherently more robust than analog circuits. Analog parts are often very sensitive to small variations and could be a bottleneck for circuit lifetime. Therefore, this work focuses on design for reliability of analog ICs. In the next section, nm CMOS reliability is discussed in more detail.

1.4 Reliability in Nanometer CMOS

The exponential increase in the gate leakage current when scaling the gate oxide thickness of CMOS transistors, forced device engineers to introduce gate materials with a higher dielectric constant compared to traditional SiO_2 or SiON gate dielectrics. This allows further increase of the gate oxide capacitance while keeping the physical gate thickness sufficiently large (Degraeve et al. 2008). Unfortunately, the introduction of new materials and devices, combined with the further reduction of the lateral transistor dimensions, reduces circuit reliability. Also, more demanding mission profiles (e.g. higher operational temperatures, high currents and extreme lifetimes) and increasing constraints on time and money further increase the required levels of circuit reliability. As a result, device and interconnect reliability has become a major focus in the ITRS guidelines (International Technology Roadmap for Semiconductors 2011) (see Fig. 1.3). Further, as shown in Fig. 1.4, the increasing amount of reliability problems is also reflected in the number of publications on the subject, which has grown exponentially over the last three decades. Below, the major lifetime-limiting scaling factors are explained briefly. Details about each individual unreliability effect will be given in Chap. 2.

1.4.1 Reduction of the Effective Oxide Thickness

In the search for suitable high-k dielectrics, most research currently focuses on HfO-based or TiN-based materials. However, none of these dielectrics is compatible with Si. This incompatibility is solved by maintaining a very thin SiO_2 or SiON interfacial layer (IL) between the silicon substrate and the high-k material. Figure 1.5 depicts a schematic representation of a traditional 90 nm CMOS stack and a modern 32 nm

Table PIDS1 Process Integration Difficult Challenges	
Near-Term 2011-2018	*Summary of Issues*
1. Scaling Si CMOS	Scaling planar bulk CMOS Implementation of fully depleted SOI and multi-gate (MG) structures Controlling source/drain series resistance within tolerable limits Further scaling of EOT with higher κ materials (κ > 30) Threshold voltage tuning and control with metal gate and high-κ stack Inducing adequate strain in new structures
2. Implementation of high-mobility CMOS channel materials	Basic issues same as Si devices listed above High-κ gate dielectrics and interface states (D_{it}) control CMOS (*n*- and *p*-channel) solution with monolithic material integration Epitaxy of lattice-mismatched materials on Si substrate Process complexity and compatibility with significant thermal budget limitations
3. Scaling of DRAM and SRAM	DRAM— Adequate storage capacitance with reduced feature size; implementing high-κ dielectrics Low leakage in access transistor and storage capacitor; implementing buried gate type/saddle fin type FET Low resistance for bit- and word-lines to ensure desired speed Improve bit density and lower production cost in driving toward $4F^2$ cell size SRAM— Maintain adequate noise margin and control key instabilities and soft-error rate Difficult lithography and etch issues
4. Scaling high-density non-volatile memory	Endurance, noise margin, and reliability requirements Multi-level at < 20 nm nodes and 4-bit/cell MLC Non-scalability of tunnel dielectric and interpoly dielectric in flash memory – difficulty of maintaining high gate coupling ratio for floating-gate flash Few electron storage and word line breakdown voltage limitations Cost of multi-patterning lithography Implement 3-D NAND flash cost effectively Solve memory latency gap in systems
5. Reliability due to material, process, and structural changes, and novel applications.	**TDDB, NBTI, PBTI, HCI, RTN in scaled and non-planar devices** **Electromigration and stress voiding in scaled interconnects** **Increasing statistical variation of intrinsic failure mechanisms in scaled and non-planar devices** **3-D interconnect reliability challenges** **Reduced reliability margins drive need for improved understanding of reliability at circuit level** **Reliability of embedded electronics in extreme or critical environments (medical, automotive, grid...)**

Fig. 1.3 Excerpt from the 2011 ITRS guideline on process integration, devices and structures. Reliability (indicated with *bold line*) is considered as a major problem in the near future

high-k metal-gate (HKMG) stack. For a transistor in inversion, the electric field E_{SiO_2} over the SiO_2 layer in each stack can be written as:

$$E_{SiO_2} = \frac{V_{GS} - V_{TH}}{EOT} \tag{1.1}$$

with V_{TH} the threshold voltage and EOT the effective oxide thickness:

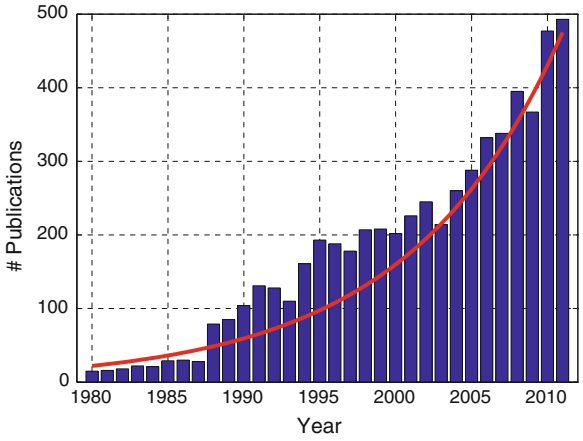


Fig. 1.4 Evolution of the number of publications with keywords 'transistor reliability' as listed by IEEE Xplore

$$\mathrm{EOT}_{90\ \mathrm{nm}} = t_{\mathrm{SiO_2}} \tag{1.2}$$

$$\mathrm{EOT}_{32\ \mathrm{nm}} = t_{\mathrm{IL}} + \frac{\varepsilon_{\mathrm{SiO_2}}}{\varepsilon_{\mathrm{HK}}} t_{\mathrm{HK}} \tag{1.3}$$

with $t_{\mathrm{SiO_2}}$ the thickness of the SiO_2 oxide in a 90 nm technology (typically $t_{\mathrm{SiO_2}} = 2.0 - 2.4\,\mathrm{nm}$), t_{IL} the thickness of the SiO_2 interfacial layer in the 32 nm technology (typically $t_{\mathrm{IL}} = 0.5 - 1\,\mathrm{nm}$) and t_{HK} the thickness of the high-k layer in the 32 nm technology (typically $t_{\mathrm{HK}} = 2 - 4\,\mathrm{nm}$). $\varepsilon_{\mathrm{SiO_2}}$ and $\varepsilon_{\mathrm{HK}}$ are the dielectric constants for SiO_2 ($\varepsilon_{\mathrm{SiO_2}} \approx 3.9$) and the high-k dielectric respectively ($\varepsilon_{\mathrm{HK}} \approx 30$). As a result, $\mathrm{EOT}_{32\ \mathrm{nm}}$ is smaller than $\mathrm{EOT}_{90\ \mathrm{nm}}$, resulting in a larger electric field over the SiO_2-interfacial layer of a HKMG technology compared to the electric field over the SiO_2 oxide in a traditional CMOS technology (i.e. $E_{\mathrm{SiO_2,\ 32\ nm}} > E_{\mathrm{SiO_2,\ 90\ nm}}$) (Degraeve et al. 2008). Since most transistor degradation effects depend exponentially on this electric field, the introduction of high-k materials further reduces the maximum operating voltage to guarantee reliable circuit operation (Degraeve et al. 2008).

1.4.2 Introduction of New Materials and Devices

Transistor unreliability effects such as negative bias temperature instability (NBTI), temperature dependent dielectric breakdown (TDDB) and hot carrier injection (HCI) were in older SiO_2 or SiON based technologies (i.e. $\geq$65 nm) considered as the most important aging effects. Both the NBTI and HCI effect generates traps at the substrate/dielectric interface (more details will be given in Chaps. 2 and 3). These traps affect transistor parameters such as the threshold voltage V_{TH}. With the introduction

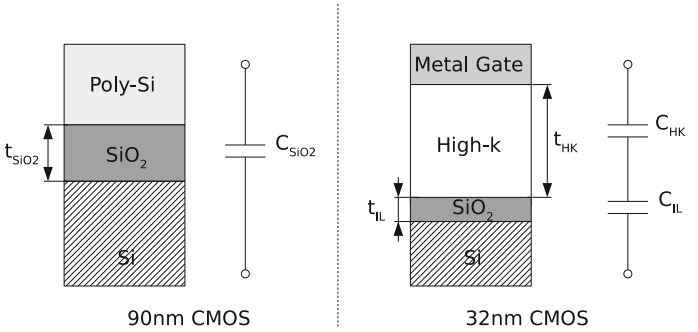


Fig. 1.5 Schematic representation of a traditional 90 nm CMOS SiO_2-based stack (on the *left*) and a 32 nm CMOS high-k metal-gate HfO-based stack (on the *right*)

of high-k materials and new device architecture such as FinFETs, a thin SiO_2 or SiON interfacial layer has however been maintained (see Fig. 1.5). Consequently, the substrate/dielectric interface does not change and NBTI and HCI remain a problem in HKMG technologies (Degraeve et al. 2008). Further, research has indicated the interfacial layer to be the major factor controlling breakdown in HKMG technologies (Bersuker et al. 2010). Therefore, models and principles previously developed to characterize breakdown in older technologies still apply in high-k technologies.

The PBTI effect, which is negligible in SiO_2 or SiON based technologies, has been found to become a lot worse in high-k technologies (Cho et al. 2010). Existing transistor failure mechanisms thus remain and even become worse with the introduction of high-k dielectrics in advanced nanometer CMOS nodes.

1.4.3 Atomic-Scale Dimensions

BTI and HCI effects in large micrometer-sized transistors are typically considered deterministic (see Chap. 3). The application of a given voltage stress on matched transistors therefore results in an identical shift of the transistor parameters (except for a statistical mismatch due to process variations). Scaling transistors down to nanometer dimensions, however, gradually has changed these deterministic effects into stochastically distributed failure mechanisms (Kaczer et al. 2010). At device level this results in a time-dependent shift of the transistor parameters (i.e. $\Delta V_{\mathrm{TH}} = \mathrm{f}(t)$) augmented with a time-dependent increase of the standard deviation on these parameters (i.e. $\sigma(V_{\mathrm{TH}}) = \mathrm{g}(t)$). After some time, initially matched transistors processed in ultra-scaled nanometer CMOS technologies can therefore cause circuit failure resulting from increased time-dependent transistor mismatch. Also, electromigration in copper lines becomes worse due to a reduction of the cross sectional diameter of the interconnection wires with scaling (Zhang et al. 2010; International Technology Roadmap for Semiconductors 2011).

1.4.4 Mission Profiles

Circuit environmental conditions and required performance tend to be stretched further with each technology generation. Sensor applications for automotive products, for example, need to function properly in temperatures exceeding 200 °C and applications such as base stations and solar cells must function reliably and almost continuously over a period of tens of years (International Technology Roadmap for Semiconductors 2011).

1.4.5 Time and Money Constraints

The constraints on time and money are always increasing. This trends is combined with possible major technology changes and therefore poses a real challenge for reliability engineers to still guarantee reliable product operation. The speed of introduction of new materials and devices reduces the capability to build up knowledge on new failure mechanisms while, at the same time, failure rate requirements become more and more demanding (International Technology Roadmap for Semiconductors 2011). The probability of an unrecognized failure mechanism to make it into an end product is therefore increasing.

1.5 Design for Reliability

The previous sections have discussed the ubiquitous presence of electronic systems in our modern society. Unfortunately, the device lifetime of the nm CMOS technology that enables this evolution reduces with each next technology generation. Some of these reliability problems can and are being solved at a device level. Already in the early 1980s, for example, alternative MOSFET structures with a graded drain junction or an offset gate structure (i.e. a lightly doped drain or LDD) to reduce the hot carrier degradation effects have been proposed (Takeda et al. 1982). A technology-based solution is however not always possible, especially since the focus of device engineers is typically on developing smaller and faster transistors with lower power consumption.

To cope with reliability problems during the design of a new product, a design for reliability (DFR) flow is used. Such a flow encompasses the entire set of tools supporting product and process design and ensuring that customer expectations for reliability are fully met throughout the life of the product. DFR relies on an array of reliability engineering tools along with a proper understanding of when and how to use these tools throughout the design cycle. DFR is of utmost importance to guarantee low warranty costs and high customer satisfaction. The technical problems with Microsoft's Xbox 360, in the early months after the console was introduced,

illustrate this perfectly (Xbox 360 Technical Problems 2012). The problem, which has never been explained officially by Microsoft, caused a general hardware failure of the console and resulted in over a billion dollars in warranties to be paid. Understanding potential reliability issues, knowing how to identify them and being able to alleviate the problem is becoming more important with the increasing complexity of systems and its interactions.

The push for a more structured approach to guarantee product reliability is similar to the strive towards high-quality products in the 1980s. The latter led to successful processes such design for six sigma (Yang and El-Haik 2003). There is however a fundamental difference between quality and reliability. Quality control assures that each product sample works as designed right after production. Time-dependent effects are rarely taken into account here. Reliability, on the other hand, is about guaranteeing with a high probability that the product will perform its intended function without failure for a designated period of time and under specified conditions.

In order to control product reliability throughout the design process, Reliasoft proposed a number of key activities (Design for Reliability: Overview of the Process and Applicable Techniques 2012) define, identify, analyze and assess, quantify and improve, validate and monitor and control. The stages are depicted in Fig. 1.6 and typical tools and methods used in each stage are also given. The different activities are valid for any industrial production process, but here they are applied to the design and production of ICs. In the next sections, the different stages are briefly explained.

1.5.1 Define

The purpose of this stage is to clearly and quantitatively define the reliability requirements and goals for a product, as well as the environmental and usage conditions. In industry, reliability specifications can come from various sources such as through contracts with customers, based on safety considerations, to remain competitive or to comply with a particular standard. The latter is particularly important in the automotive industry where the automotive electronics council (AEC) has established a series of standards for integrated circuits: the AEC-Q100 (stress qualification for ICs) and the AEC-Q200 (stress test qualification for passive components). These standards are mainly used in the US, but approved by many automotive electronics companies around the world. In Europe, the main standard used is the ISO 16750 series. Other commonly used IC reliability standards, also used for consumer or industrial applications, are MIL-STD 883, MIL-STD 750, JESD 47, JP001.01, JESD22A-108, JESD85, JESD74, JEP122 and EIAJ 4701-100. Table 1.1 summarizes the most important reliability-related requirements that are typically applied in different markets. As expected, specifications for the consumer market are clearly less demanding compared to industrial or automotive applications.

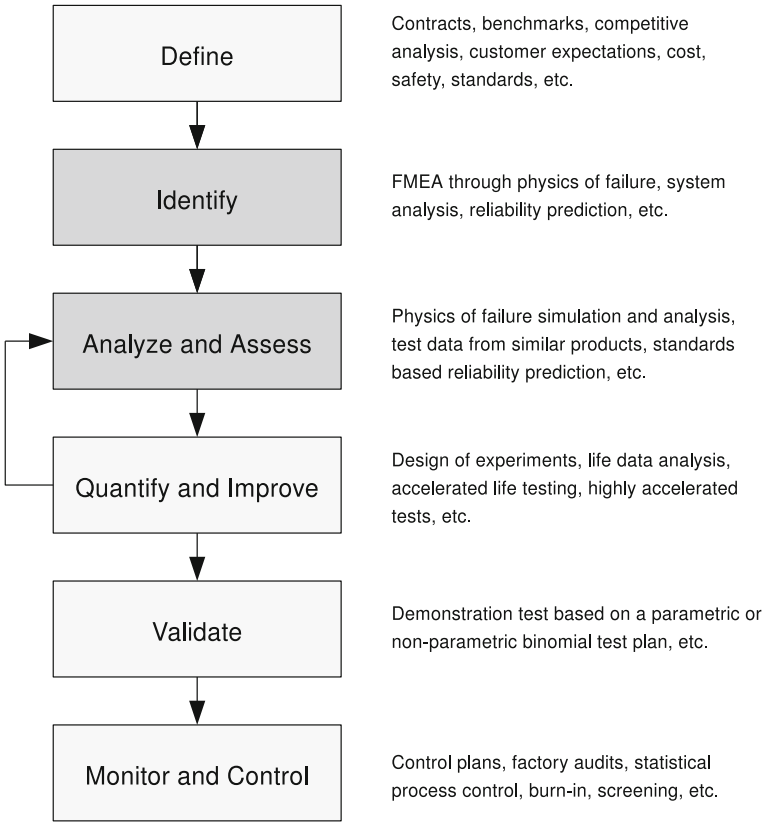


Fig. 1.6 A general design for reliability (DFR) flow. Typical tools and methods used in each stage are also indicated. This work focuses on the design of reliable analog ICs therefore and studies an implementation of the second and third step of the above DFR flow (Design for Reliability: Overview of the Process and Applicable Techniques 2012)

1.5.2 Identify

In this stage, possible reliability threats are identified. A failure mode effect analysis (FMEA) strategy can be used as a tool to quantify the risk associated with different failure effects. The failure rate or frequency with which an electronic system fails is typically represented using the so-called bathtub curve (see Fig. 1.7). Circuits that fail right after production are referred to as infant mortality failures. In IC design, these failures primarily result from process errors such as oxide defects, mask defects, contamination, bonding issues, solder defects, etc. Infant mortality failures typically happen within the duration of the product warranty and therefore have to be limited. During the useful life of the product, random defects and soft errors can result in an occasional circuit failure. Then, the product starts to wear out due to aging effects and

Table 1.1 Typical IC reliability requirements

	Consumer	Industrial	Automotive
Operating temperature	−5/0 °C −40/65 °C	−10 °C −70 °C	−40 °C −85/155 °C
Lifetime (years)	1–5	5–10	15–20
Tolerated failure rates	<10 %	<1 %	Target: zero failures
Humidity (%)	30−85	15−90	0−100
Condensation	Low	Medium	High
Temperature cycling	Low	Medium	High
Altitude (km)	7–10	10–12	12–15
Vibration	Low	Medium	High
Shock/bump	Low	Medium	High
Sand/dust	Low	Medium	High

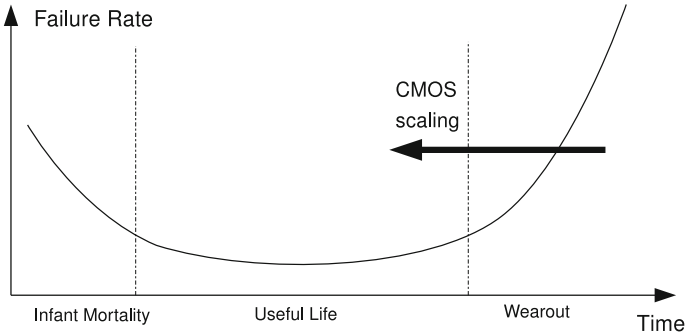


Fig. 1.7 The *bathtub* curve represents the number of failures or the failure rate of a product over time

eventually every IC fails. The focus of this work is on the wearout stage, which sets in earlier with every new CMOS generation (also see Sect. 1.4) (Franco et al. 2010).

1.5.3 Analyze and Assess

During the design phase, the product lifetime is estimated and expressed as the mean time to failure (MTTF). The MTTF also helps to compare different design concepts and can be used to identify design margins and to assess failure-resilient design strategies. To estimate the product lifetime (time to wearout), industry mainly uses accelerated stress tests on individual devices. Further, failure criteria are often chosen arbitrarily (e.g. a 10 % shift in I_D) and the impact of device failure on the circuit level is typically not considered (Groeseneken et al. 2010). Circuit MTTF figures are therefore hard to estimate and rarely circuit specific. Designers are forced to use large design margins, ultimately limiting circuit performance or costing a large

area or power overhead, and even then uncertainty about the lifetime of a circuit remains. To solve these problems, reliability assessment needs to be done at design time, requiring:

- Accurate transistor compact models for all important transistor unreliability effects.
- Efficient circuit simulation techniques to i) analyze the reliability of a circuit and ii) to identify circuit reliability weak spots.

These are discussed further in this work for the aging-related effects in CMOS circuits.

1.5.4 Quantify, Improve and Validate

During this stage, measurement results are used to verify the simulation results obtained from previous stages. This testing is typically done on prototype circuits using accelerated life tests (ALT). The results are analyzed, root causes for failure are identified and if necessary design changes are done and the tests are repeated. ALT was first developed in the 1960s to cope with the high-reliability requirements of the US space program. Old methods for product validation took too long times and were therefore no longer adequate to meet the short time-to-market demands. Therefore, a way to quickly identify product defects and to solve them effectively is the key issue for major manufacturers in the world. Today, HALT (highly accelerated life test) and HAST (highly accelerated stress test) are used to solve the problem. Starting in 1990, major manufacturers including HP, Dell, Cisco, Nortel, Tektronix and Motorola successfully employed a HALT approach to speed up the identification of design and production defects, with improvements being capable of lowering the cost of the warranty period, enhancing the reliability of the product and shortening the time to market. The failure modes identified by HALT, together with relevant information, can also be used as input for developing new products. HALT is done with a chamber, as depicted in Fig. 1.8, enabling gradually increased voltage and temperature stressing such as high and low thermal cycling, combined stress, power switch cycling and voltage and frequency margin test. If a reliability problem is detected, design changes are needed and the activities described above are repeated until the product is considered to be acceptable. Further, extra variations introduced by the manufacturing process also need to be taken into account and if necessary design modifications are required to account for this. After this stage, the product is ready for volume production.

1.5.5 Monitor and Control

Once a product is in production, the process is monitored to assure that process variations are kept under control and that reliability is still guaranteed. Burn-in techniques

Fig. 1.8 A highly accelerated life test (HALT) chamber enables the reliability testing of electronic circuits at elevated voltages and temperatures

are used to prevent infant mortality failures due to manufacturing-related problems (Vassighi et al. 2004). Further, continuous monitoring and field data analysis can help to observe the behavior of the circuit in actual use conditions and to gain knowledge for further improvements.

1.6 Conclusions

This introductory chapter has discussed the importance of long- and medium-term reliability in various application domains. Circuits designed in nanometer CMOS technologies suffer more than ever from transistor aging effects. Therefore a design for reliability strategy is needed. The different stages of such a DFR flow have been discussed and the current industrial approach has been reviewed. Early life failures due to defects related to the production process are eliminated with careful process control and circuit burn-in testing. Wearout of ICs, however, is typically only evaluated at the device level and a lack of circuit-level evaluation and assessment results in the use of large design guardbands, costing a large area and/or power overhead. The focus of this work is on the development of transistor compact models and simulation methods for analog circuit lifetime analysis in nm CMOS. Further, the goal of this work is also to assess the impact of these aging effects on analog circuits.

Chapter 2
CMOS Reliability Overview

2.1 Introduction

For over four decades, scientists have been scaling devices to increasingly smaller feature sizes (Lewyn et al. 2009; International technology roadmap for semiconductors 2011). This trend is driven by a seemingly unending demand for ever-better performance and by fierce global competition. The steady CMOS technology downscaling is needed to meet requirements on speed, complexity, circuit density, power consumption and ultimately cost required by many advanced applications. However, going to these ultra-scaled CMOS devices also brings some drawbacks.

This chapter discusses the most important effects designers have to deal with in order to manufacture reliable integrated circuits in nanometer CMOS processes. The intent of this chapter is not to give an in-depth description of the physics behind each failure mechanism, but to provide the reader with a basic understanding of the most important unreliability effects and how these effects evolve with technology. First, Sect. 2.2 briefly outlines how various unreliability effects came into play in the course of history. Next, Sect. 2.3 reviews the most important spatial unreliability effects in modern CMOS technologies. These effects are related to process variations and are visible right after production. A difference is made between systematic and random effects. Time-dependent unreliability effects are then discussed in Sect. 2.4. These effects are divided into aging and transient effects.

2.2 The Origin of CMOS Unreliability

Device reliability was first studied in the early sixties, when increasingly complex integrated systems were developed and fabricated. Conferences such as the first international reliability physics symposium (IRPS 1962, Chicago) were the first attempts to bring engineers and scientists together from all over the world to study the physics behind various failure effects (Physics of failure in electronics 1962).

E. Maricau and G. Gielen, *Analog IC Reliability in Nanometer CMOS*,
Analog Circuits and Signal Processing, DOI: 10.1007/978-1-4614-6163-0_2,

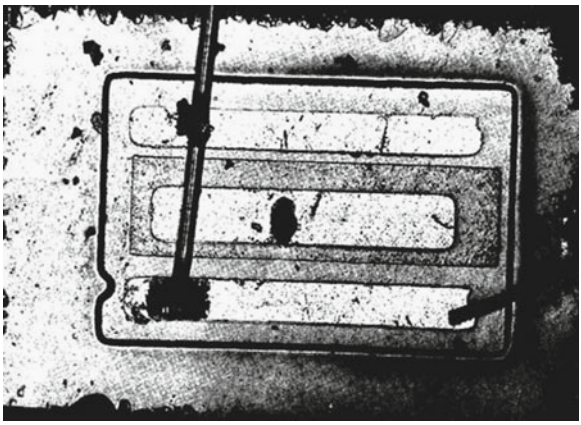


Fig. 2.1 Photomicrograph of an early silicon mesa transistor on which the emitter bond has separated due to 'purple plague' (Phillips et al. 1962). This phenomenon, also known as 'purple death', was an important reliability problem in the late 1960s and the early 1970s. An intermetallic reaction between the golden bond wires and the aluminum bond pads formed a brittle bright purple compound of $AuAl_2$ which led to the creation of voids in the metal lattice

During the 1970s, effects such as corrosion, bonding issues (e.g. the 'purple plague' as depicted in Fig. 2.1) and ionic contamination were the most common causes of circuit failure. All these issues were however related to the way how integrated circuits were packaged and mounted on a printed circuit board (PCB). Only in the late seventies and early eighties the first real integrated circuit reliability issues became visible. Oxide thickness scaling increased the gate-oxide electric field, and transistor wearout effects such as hot carrier injection (HCI) started to affect device performance within the lifetime of a circuit (Takeda et al. 1983; Hu et al. 1985). Initially, the application of an arbitrary voltage stress resulted in an identical parameter shift for matched devices. Therefore these temporal unreliability effects were at first considered as deterministic. However, when the oxide dielectric reached atomic-scale dimensions, this resulted in the first stochastic temporal unreliability effect: time-dependent dielectric breakdown (Solomon 1977). Further, matched devices were, in the early eighties, considered identical in terms of electrical performance. In the second half of that decade, however, when device dimensions entered the nanometer scale, stochastic errors and variations at atomic level became apparent at device level and sensitive analog circuits were the first to suffer from process variability effects (Lakshmikumar et al. 1986; Pelgrom et al. 1989). Device mismatch became a big issue (especially analog) designers had to deal with in order to guarantee good accuracy and high yield.

To overcome scaling limitations of devices fabricated in ultra-scaled CMOS processes, changes in device structures, processing materials and processing conditions have been introduced. These changes have drastically increased the complexity of nanometer CMOS technologies. Examples of these new techniques include:

Table 2.1 Evolution of nanometer CMOS characteristics

Year	L_g (nm)	V_{DD} (V)	V_{TH} (V)	EOT (nm)	E_{ox} (MV/cm)
1995	350	3.3	0.58–0.70	10.0–12.0	2.17–2.72
1998	250	1.8–2.5	0.47–0.52	6.0–7.0	1.83–3.38
2003	180	1.8	0.39–0.43	4.5–5.5	2.49–3.13
2001	130	1.2	0.35–0.40	3.5–4.0	2.00–2.43
2004	90	1.0–1.2	0.25–0.40	1.6–3.0	2.00–5.93
2007	65	1.0–1.2	0.20–0.35	1.5–2.0	3.25–6.66
2009	45	1.0–1.1	0.20–0.35	1.0–1.4	4.64–9.00
2011	32	0.9–1.0	0.20–0.35	0.8–1.1	5.00–10.0

(Iwai 1999; Bult 2000; Bravaix et al. 2009; Wu et al. 2009; Europractice 2012; International technology roadmap for semiconductors 2011)

strained silicon channels to increase the transistor drive current, the introduction of high-k oxides and metal gates to allow further gate oxide scaling combined with reduced gate leakage, and Cu-interconnect with low-k dielectrics to ensure lower RC-delays (Horstmann et al. 2009; International technology roadmap for semiconductors 2011). However, the introduction of these new materials also increased the impact of already existing but before then unimportant aging effects, such as electromigration (EM) and negative bias temperature instability (NBTI), and even created new problems such as positive bias temperature instability (PBTI) (Lewyn et al. 2009; Groeseneken et al. 2010). Table 2.1 gives an overview of typical technology parameters for the most recent CMOS nodes. The table clearly shows how the average oxide electric field increases with each new technology node, aggravating all transistor wearout effects. All of these phenomena can have a large impact on the reliability of a circuit, right after production or during its operational lifetime. Therefore, a good understanding of the impact of each effect on the electrical behavior of a single transistor and eventually on the performance of an entire circuit is mandatory.

Figure 2.2 illustrates how nanometer CMOS reliability issues can be categorized into spatial and temporal unreliability effects. Spatial unreliability effects are immediately visible right after production and are fixed in time. Spatial unreliability effects can be random (e.g. random dopant fluctuations (RDF), line edge roughness (LER), etc) or systematic (e.g. gradient effects, etc.). The effects depend on the circuit layout, the neighboring environment, process conditions and the impact the geometry and structure of the circuit and can lead to yield loss. This yield loss can be functional or parametric, i.e. resulting in malfunctioning circuits or circuits with degraded performance respectively. Temporal unreliability effects, on the other hand, are time-varying and change depending on operating conditions such as the operating voltage, temperature, switching activity, presence and activity of neighboring circuits. A difference is made between wearout or aging effects (e.g. hot carrier injection (HCI), NBTI, etc.) and transient effects (e.g. electromagnetic interference (EMI), single event upsets (SEU), etc.). In the following sections, these effects are discussed in more detail.

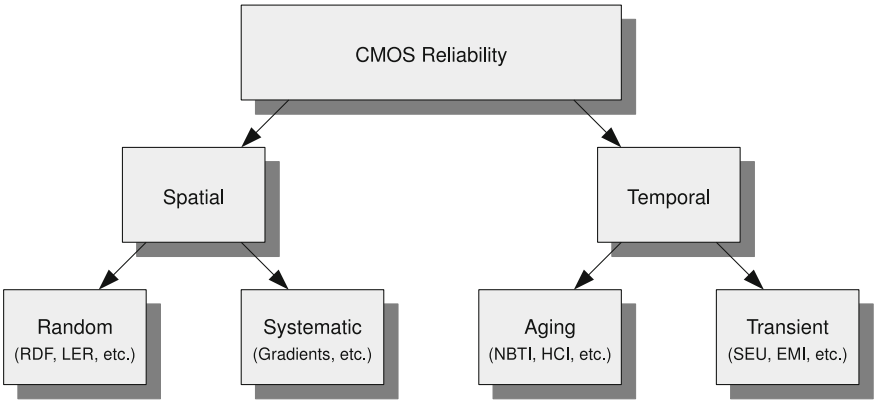


Fig. 2.2 A CMOS circuit can fail from spatial or temporal unreliability effects. The former are visible right after production and can be random or systematic. The latter become a potential problem during the operational lifetime of the circuit and present themselves as an aging effect or a transient effect

2.3 Spatial Unreliability

Spatial unreliability or process variability is an increasing problem in nanometer CMOS IC production. The problem results from the increasing complexity needed to fabricate nanometer CMOS devices, combined with the scaling towards atomistic device dimensions (<180 nm CMOS). Typically, parametric yield is used as a metric to express the impact of these effects on the performance of the circuit right after production.[1] A high yield implies low spatial unreliability.

Two major sources of process variability are distinguished: local or intradie and global or interdie effects. Local variability results in parametric variations of identically designed transistors across a short distance, typically within the same circuit. This is also referred to as device mismatch. Global variability refers to variations between devices that are separated by a long distance or that are fabricated at a different time. Typically global variability is variability from die to die, wafer to wafer or lot to lot. Global variability causes a shift in the mean value of design parameters such as channel length or doping density. Since most spatial unreliability problems result from local variability effects, global variability is not discussed here. Local variability originates from systematic and random reliability effects. The former includes variability caused by optical proximity correction, layout-induced strain and well-proximity effects. The latter includes random dopant fluctuation (RDF) effects, line edge and width roughness (LER and LWR), fixed charges in the gate dielectric and interface roughness. Systematic variability is typically addressed through

[1] Parametric circuit failures are related to process variations and are circuits that do function but with a performance outside the desired range. Catastrophic circuit failures result from process errors or defects and are described by the functional yield. The latter are not covered in this work.

careful layout design, compensating circuit techniques and with advanced manufacturing flows. Solving random variability issues, on the other hand, requires innovative process and design techniques and accurate device models. For technology generations below 90 nm CMOS, the impact of random variability is becoming increasingly important (Lewyn et al. 2009). Both the systematic and random variability effects are discussed in more detail in the next sections.

2.3.1 Systematic Effects

While most device-related sources of spatial unreliability are random, a large fraction of the variation of the interconnect is a function of layout characteristics (i.e. design dependent). These sources of variability have a large systematic component. With the aggressive scaling to smaller feature sizes, this component has become larger primarily due to resolution limitations. The inability to scale the wavelength of the light source for lithography has led to an increase of systematic variations, especially in circuit areas with high interconnect and device density (Agarwal and Nassif 2007). To mitigate these problems, a lot of research has gone into more advanced manufacturing flows such as double-patterning technologies (DPT), optical-proximity correction (OPC), extreme ultraviolet lithography (EUVL) and into design techniques such as the use of extremely regular circuit layout (Strojwas 2011).

2.3.2 Random Effects

Random spatial unreliability results from physical phenomena such as random dopant effects, line edge and width roughness, fixed charges in the gate dielectric and oxide thickness variation resulting from interface roughness (Agarwal and Nassif 2007). Random effects typically affect the mismatch between closely spaced identically designed devices. At device level these effects mainly result in variations of the gate length (L), the threshold voltage (V_{TH}) and the current factor (β) (Zhao et al. 2007), and they can to first order be modeled with Pelgrom's model (Pelgrom et al. 1989):

$$\sigma(\delta V_{\mathrm{TH}}) \approx \frac{A_{\mathrm{VTH}}}{\sqrt{WL}} \tag{2.1}$$

where $\sigma(\delta V_{\mathrm{TH}})$ is the standard deviation on the threshold voltage mismatch between two identically sized transistors, WL is the size of the active area of one transistor and A_{VTH} is a technology-dependent constant commonly expressed in $\mathrm{mV}\,\mu\mathrm{m}$.[2]

[2] Pelgrom's model expresses the standard deviation on the *difference* between the threshold voltages of two matched transistors. The standard deviation on the threshold voltage of a single transistor can be found by dividing A_{VTH} by $\sqrt{2}$.

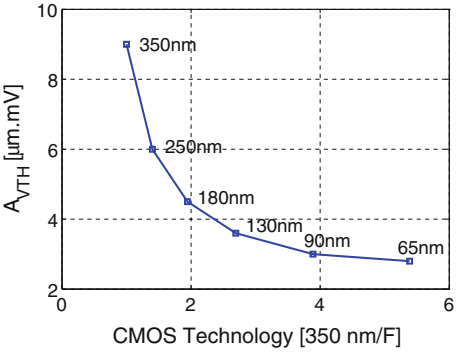


Fig. 2.3 Measured A_{VTH} values for minimum-length pMOS devices as a function of 350 nm over the minimum feature size F. Extrapolating the curve, significant improvements are not expected beyond the 65 nm technology node. Data taken from Lewyn et al. (2009)

The actual relationship is more complex than (2.1) (Hong et al. 2011). Currently, IC foundries supply Monte-Carlo (MC) simulation models to accurately simulate the impact of process variations on transistor and circuit performance. Pelgrom's formula is however still used by designers for initial circuit design. It is therefore interesting to look at A_{VTH} trends in advanced technology nodes (see Fig. 2.3). From Fig. 2.3 it is clear how A_{VTH} does not improve much beyond the 90 nm technology node. Where for older technologies (>180 nm) transistor matching was primarily determined by lithographic accuracy that scaled well with technology, other factors that do not scale well are now taking over (Lewyn et al. 2009). These effects are discussed in the following sections.

Random Dopant Fluctuations

Variations of device parameters such as the transistor V_{TH} partly result from fluctuations in the amount and location of dopant atoms in the transistor channel (Takeuchi et al. 2007; Kuhn et al. 2008). Since the number of dopant atoms in the channel of scaled transistors is always decreasing, the impact of the variation associated with the atoms increases. Figure 2.4 illustrates the decreasing average number of dopant atoms as a function of technology. Note how the number of dopants decreases by almost three orders of magnitude when going from a 1 µm (with around 1e4 dopant atoms) to a 32 nm technology (with less than 100 dopant atoms). Random dopant fluctuations (RDF) are assumed to be the major contributor to device mismatch of identically designed devices. For example, in (Kuhn 2007) it is shown how the simulated RDF is responsible for around 65 % of the total NMOS V_{TH} variation of a 65 nm CMOS technology. Similar results were obtained for a 45 nm PMOS transistor, where RDF was responsible for 60 % of the $\sigma(V_{TH})$. The effect of RDF on the transistor threshold variation is frequently represented by (Stolk et al. 1998):

$$\sigma(V_{TH}) \propto \frac{t_{ox}}{\varepsilon_{ox}} \frac{\sqrt[4]{N}}{\sqrt{W_{eff} L_{eff}}} \tag{2.2}$$

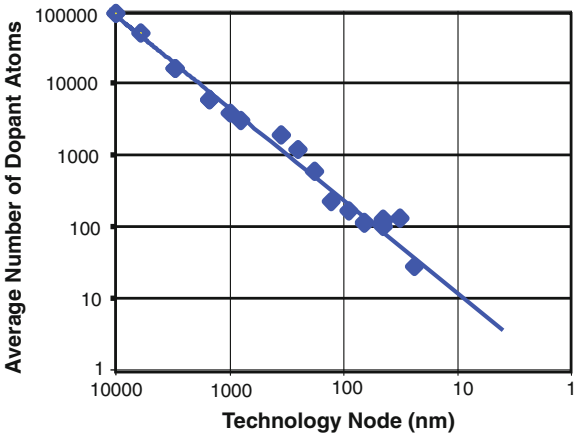


Fig. 2.4 Average number of dopant atoms in the channel of a transistor as a function of the technology node (Kuhn et al. 2008)

with t_{ox} the gate oxide thickness, ε_{ox} the oxide permittivity, N the number of channel dopants and W_{eff} and L_{eff} the transistor effective width and length. Equation (2.2) shows that V_{TH} matching improves with technology scaling (both t_{ox}/ε_{ox} and N reduce with smaller feature sizes). However, the device area ($W_{eff} L_{eff}$) also decreases with each new technology generation. Therefore the net result of RDF is a significant increase in process variability for scaled CMOS technologies.

Line Edge/Width Roughness

Line edge and line width roughness (LER and LWR respectively) result from subwavelength lithography. Since the 0.25 μm technology generation, the semiconductor industry has used subwavelength lithography to pattern transistors. Fabrication processes initially used the wavelength of light ($\lambda = 248$ nm) to pattern minimum feature sizes of 250 and 180 nm transistors. Then the value of λ decreased to 193 for 130 nm transistors and it has remained there ever since, even for 65 nm and smaller transistors. As shown in Fig. 2.5, this lithographic gap causes the LER and LWR effect (saha 2010). Although LER and LWR occur in both the front-end and the back-end of a CMOS process, LER and LWR in the poly-gate patterning are the primary concern (Kim et al. 2004). This results in both an increase of the subthreshold current as well as a variation of the V_{TH} (Asenov et al. 2003; Fukutome et al. 2006). Further, assuming the variations on the line edge are not correlated, $\sigma(\text{LWR}) = \sqrt{2}\sigma(\text{LER})$. Asenov et al. (2003) studied the combined effect of LER and RDF on current fluctuations. They demonstrated that these two sources of transistor variability are statistically independent. Experiments have shown that LER is on the order of 5 nm and does not scale with technology. Also, LER has a much stronger channel length dependence compared to RDF. LER is expected to replace

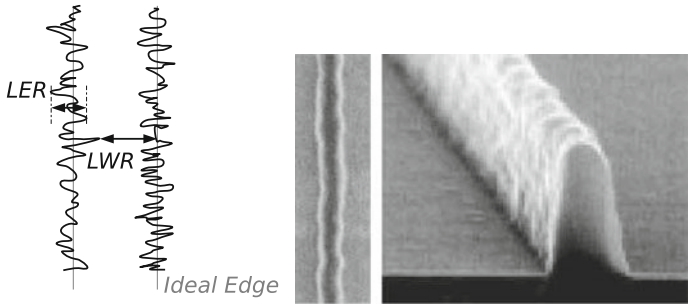


Fig. 2.5 Line edge roughness and line width roughness are a major source of process variations and result in an increase of the subthreshold current and variations on the V_{TH}. Photo: (Mack 2006)

RDF as the dominant source of transistor mismatch as device scaling continues. The transitional channel length is around 45 nm and depends on the actual device architecture and on the lithographic process used to fabricate the devices.

Gate Dielectric Variations

The gate dielectric can suffer from various non-idealities and defects such as variations in the oxide thickness, fixed charges in the oxide and interface traps. These physical effects result in parametric variations in the drive current, the gate tunneling current and/or the V_{TH} (Kuhn et al. 2008; Saha 2010).

Asenov et al. (2003) have shown that V_{TH} fluctuations induced by local oxide thickness variations become comparable to voltage fluctuations introduced by RDF for sub-30 nm CMOS technologies. Moreover, the variability due to oxide thickness fluctuations is statistically independent from the V_{TH} variations introduced by RDF. HKMG devices allow larger physical oxide thicknesses, but still suffer from large oxide thickness variations due to the roughness of the interface between the silicon and the high-k layer and between the high-k layer and the metal gate (saha 2010).

High-k gate dielectrics also suffer from fixed charges in the gate-oxide layer. These charges can affect the carrier mobility and the V_{TH}. As a consequence, variations in the location of these charges may affect the distribution of the mobility and the V_{TH} (Kaushik et al. 2006). Further, electron mobility degradation and V_{TH} instability due to fast transient charging (FTC) in interface traps is an increasing concern for high-k dielectrics (Kuhn et al. 2008).

Other Sources

Other sources of random process variability include: variations associated with patterning proximity effects such as optical proximity correction (OPC), variations associated with polish such as shallow-trench isolation and its effect on gates and

interconnections and variations associated with strain such as high-stress capping layers and embedded silicon germanium layers (Kuhn et al. 2008; Saha 2010).

2.4 Temporal Unreliability

Temporal unreliability becomes apparent after a circuit has been produced, when it is used in a certain environment, at a given temperature and workload and over a period of time. The impact of these effects on the circuit can be permanent or temporary. Aging effects cause a gradual degradation of the circuit (which does not always directly result in reduced circuit performance) and at least part of the damage is permanent. Transient effects only temporarily distort the circuit performance and the circuit performs back as before once the noise source is removed.

2.4.1 Aging Effects

Integrated circuit aging phenomena were first observed during the seventies and the eighties. At that time, research effort was mainly focused towards understanding these effects, rather than solving circuit reliability problems. In the nineties, however, circuit aging became more and more an issue due to the aggressive scaling of the device geometries and the increasing electric fields. At that time, measurements on individual transistors were used to determine circuit design margins in order to guarantee reliability. After the turn of the century, the introduction of new materials to further scale CMOS technologies introduced additional failure mechanisms and made existing aging effects more severe. This section reviews the most important integrated-circuit aging phenomena observed in sub-90 nm CMOS technologies: hot carrier injection (HCI), time-dependent dielectric breakdown (TDDB), bias temperature instability (BTI) and electromigration (EM).

Hot Carrier Injection

In general, hot carriers are particles that obtain a very high kinetic energy from being accelerated in a high electric field. These energetic carriers can be injected into 'forbidden' regions of the device, such as the gate oxide, instead of following their intended trajectory. When injected into such a region they can get trapped or cause the generation of interface states. These defects in turn lead to shifts in the electrical characteristics of the transistor such as a shift of the V_{TH}, the current factor β and the output conductance g_{o}. The degradation of integrated circuits due to hot carrier injection (HCI) first became a problem in the mid-eighties due to the continuous scaling of transistor dimensions without accompanying supply voltage reduction (Takeda et al. 1983; Tam et al. 1984; Hu et al. 1985). In the mid-nineties,

the circuit operating voltage was dropped to reduce power consumption and graded drain junctions were introduced to solve reliability problems. Hence, HCI became less of an issue. Also, measurements on advanced HKMG CMOS transistors revealed how high-k stacks appear to be more resilient to HCI stress than SiO_2 stacks (Amat et al. 2009). Nevertheless, HCI can still be a problem since supply voltage scaling is slowing down in recent years because of the non-scalability of the subthreshold slope (Wang et al. 2007; Maricau et al. 2008; Bravaix et al. 2009). HCI is primarily a problem in nMOS devices (Lunenborg 1996; Parthasarathy 2006). Nevertheless, although pMOS devices are less sensitive to HCI, the effect can enhance other aging effects such as negative bias temperature instability (NBTI) (Parthasarathy 2006).

As illustrated in Fig. 2.6, four different hot carrier injection mechanisms can be distinguished (Takeda et al. 1983): channel hot electron (CHE) injection, drain avalanche hot carrier (DAHC) injection, secondary generated hot electron (SGHE) injection and substrate hot electron (SHE) injection.

1. When the gate voltage is approximately equal to the drain voltage, the **channel hot electron** (**CHE**) injection effect is at its maximum. So-called 'lucky electrons' gain sufficient energy to surmount the Si/SiO_2 barrier at the drain end of the channel, without loosing energy due to collisions with atoms in the channel (see Fig. 2.6a). For low gate voltages, the field does not attract electrons to the

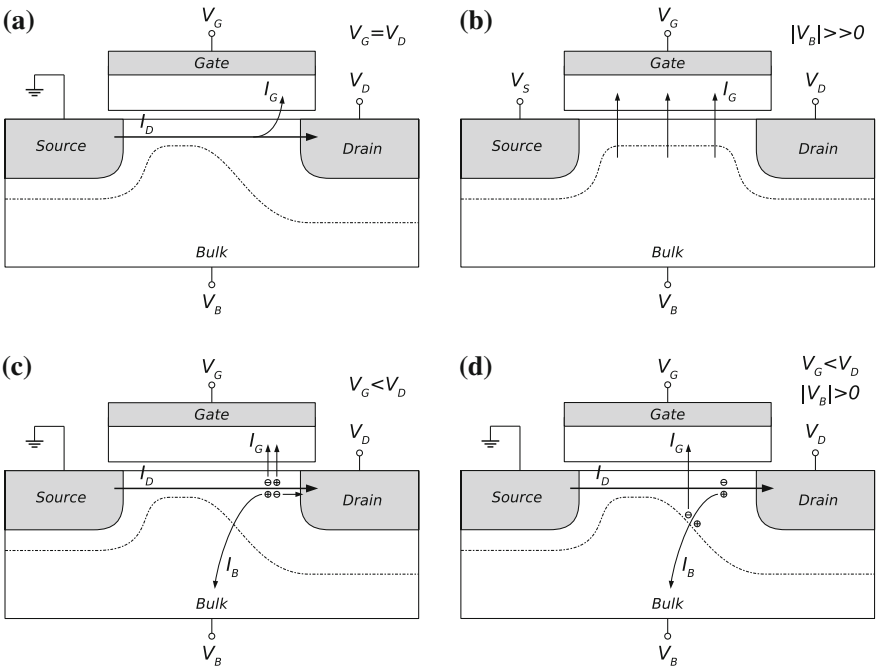


Fig. 2.6 Four different hot carrier injection mechanisms can be distinguished. **a** CHE. **b** SHE. **c** DAHC. **d** SGHE

gate electrode and for high drain voltages the electric field at the drain leads to avalanche multiplication resulting in drain avalanche hot carrier generation. As holes are much 'cooler' (i.e. heavier) than electrons, the channel hot carrier effect in nMOS devices is shown to be more significant than in pMOS devices (Hu et al. 1985).

2. **Substrate hot electron** (**SHE**) or substrate hot hole (SHH) injection is the result of a high positive or a high negative bias at the bulk of the transistor. This leads to carriers in the substrate driven to the Si/SiO_2 interface, gaining kinetic energy and potentially surmounting the energy barrier at the channel/gate-oxide interface to be injected into the oxide. In contrast to the other hot carrier generation mechanisms, this effect is uniformly distributed along the channel instead of being concentrated near the drain of the transistor (see Fig. 2.6b). This generation mechanism is especially present in circuits where stacked devices, typically implying a non-zero bulk bias, are used (e.g. current-source differential pairs and cascode circuits).
3. At stress conditions with high drain voltage and low gate voltage, electron-hole pairs can be created due to impact ionization of the channel current near the drain of the transistor. Each of these electrons and holes can then accelerate in the channel electric field and can potentially surmount the Si/SiO_2 barrier to get trapped or to create interface states. This phenomenon is known as avalanche multiplication and results in **drain avalanche hot carrier generation** (**DAHC**) (see Fig. 2.6c). Additionally, some of the generated carriers lead to a bulk current. The DAHC injection mechanism causes the most stringent device degradation because a large amount of hot electrons are injected into the gate oxide at the same time.
4. **Secondary generated hot electron injection** (**SGHE**) involves the generation of hot carriers from impact ionization with a secondary carrier that was created by an earlier impact ionization incident. This earlier generated carrier can be generated under DAHC conditions or from photons generated in the high field region near the drain (i.e. bremsstrahlung radiation). Under the influence of the field generated by the substrate's bulk bias, the first carriers are accelerated and potentially generate secondary carriers. These secondary carriers also accelerate in the bulk bias field towards the surface region where they further gain kinetic energy to overcome the surface energy barrier (see Fig. 2.6d). SGHE is observed as a rather small effect with limited contribution to the transistor degradation.

As explained above, each of the four hot carrier mechanisms occurs at different transistor operating conditions. Typically, DAHC and CHE effects are much worse than SHE and SGHE effects and therefore limit the device and circuit lifetime. For transistors with a minimum gate length of $0.35\,\mu m$, DAHC has the worst effect on transistor performance and is at its maximum when $2V_{GS} = V_{DS}$. For smaller transistor dimensions, on the other hand, CHE dominates the hot carrier degradation effect (Takeda et al. 1983).

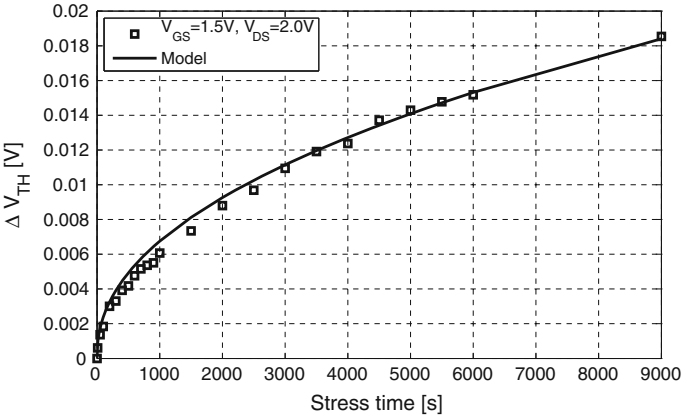


Fig. 2.7 Hot carrier injection (HCI) is typically modeled with a power law time dependence. Here the (measured and modeled) V_{TH} shift of a 65 nm nMOS transistor, stressed with $V_{GS} = 1.5\,V$ and $V_{DS} = 2.0\,V$, is depicted (Maricau et al. 2008)

HCI is typically modeled with a power law dependence on the stress time (see Fig. 2.7) (Hu et al. 1985; Kufluoglu and Ashraful Alam 2004; Maricau et al. 2008):

$$\Delta V_{TH} = A_{HCI} t^{n_{HCI}} \tag{2.3}$$

where ΔV_{TH} represents the HCI-induced V_{TH} shift and n_{HCI} is the time exponent which is typically around 0.5 (Hu et al. 1985). The trapping generation of the carriers increases exponentially with increasing oxide electric field (E_{ox}). Besides the oxide electric field and the maximum lateral electric field (E_{lat}), HCI dependence on temperature (T) and transistor length (L) has also been reported (Hu et al. 1985; Wang et al. 2007; Maricau et al. 2008):

$$A_{HCI} \propto \frac{1}{\sqrt{L}} \exp\left(\alpha_{HCI,1} E_{ox}\right) \exp\left(-\frac{\alpha_{HCI,2}}{E_{lat}}\right) \tag{2.4}$$

with $\alpha_{HCI,1}$ and $\alpha_{HCI,2}$ technology-dependent parameters. In addition to the average effect, predicted by (2.3) and (2.4), HCI also introduces an extra source of variability, due to the randomly generated traps in the gate dielectric or at the substrate/dielectric interface. Further, this effect has been shown to be more pronounced for sub-65nm technologies (Magnone et al. 2011).

Time-Dependent Dielectric Breakdown

The correct operation of a MOS transistor relies on the insulating properties of the dielectric layer below the gate electrode of the transistor (Stathis 2001). Each dielectric material has a maximum electric field it can sustain. When a larger electric

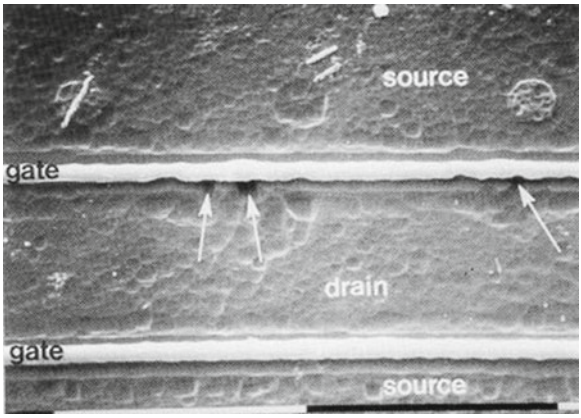


Fig. 2.8 Multiple breakdown spots at the drain junction of an nMOS transistor. Note the thermal damage to the silicon. *Source* Yazdani (2011)

field is applied, this leads to hard breakdown (HBD). HBD is an extremely local phenomenon, characterized by a loss of the gate oxide insulating properties and allowing a large gate current to flow.[3]

At lower electric fields, the insulator can wearout after some time and finally break down completely. This is called time-dependent dielectric breakdown (TDDB) (Fig. 2.8).

Prior to oxide TDDB, a degradation process of the dielectric takes place that initiates the generation of traps in random positions inside the oxide and at the interface. A stress-induced leakage current (SILC) is produced during this degradation stage (Kamohara et al. 1998; Young et al. 2012). If the dielectric degradation increases, a critical trap density is reached and BD occurs. Due to this behavior HBD is a stochastic phenomenon and can be described using a Weibull probability distribution (Wolters and van der Schoot 1985; Wu and Su 2005):

$$F(t_{\mathrm{BD}}) = 1 - \exp\left[-\left(\frac{t_{\mathrm{BD}}}{\alpha_{\mathrm{BD}}}\right)^{\beta_{\mathrm{BD}}}\right] \tag{2.5}$$

with $F(t_{BD})$ the cumulative density function for the time-to-BD. α_{BD} and β_{BD} are technology-dependent parameters.

[3] Besides TDDB, which is a time-dependent wearout effect, oxide BD can also result from electrical overstress (EOS), electrostatic discharge (ESD) or under the presence of weak spots in the oxide. EOS and ESD involve the application of a high voltage being applied across the oxide. This causes a dramatic increase of the gate current, localized heating and a meltdown of the silicon. Early life BD failures due to weak spots in the oxide are essentially similar to TDDB, but happen within the first year of the circuit operational life. This work focuses aging effects, therefore EOS, ESD and early life failures are not discussed here.

During a breakdown degradation process, different BD modes can be distinguished. Depending on the thickness of the gate oxide, one or more modes occur. The most harmful mode, the Hard-BD (HBD), provokes the complete loss of the oxide dielectric properties with gate currents in the mA range at standard operation voltages. However, HBD is in nanometer CMOS technologies only a significant reliability threat at elevated operating voltages (i.e. $V_{GS} > 1.2V$ for EOT $= 0.9$ nm) (Degraeve et al. 2008; Pae et al. 2010).

For oxide thicknesses below 5 nm (i.e. sub-180 nm CMOS), HBD can be preceded by Soft-BD (SBD). SBD can be observed as a partial loss of the dielectric properties, resulting in a small increase of the gate current and a significant increase of the gate current noise (Gielen et al. 2008). Finally, in ultra-thin oxides (approximately below 2.5 nm thickness), SBD is followed by Progressive-BD (PBD), until final HBD. PBD is detected as a slow increase of the gate current over time (see Fig. 2.9).

When looking at the impact of BD on the transistor electrical characteristics, it has been shown that the degradation process prior to BD (Martín-Martínez et al. 2007) and the BD spot location (Fernández 2007) can vary largely for transistors of the same size and therefore have a strong influence on the channel current. The transistor geometry also has a strong impact on this current. Although right after SBD a very limited effect is observed (Kaczer et al. 2004), a significant influence on the transistor characteristics is produced at longer times (Kaczer et al. 2004; Cester et al. 2004). This can be modeled as a local mobility reduction in the BD region (Cester et al. 2004). Another important aspect of gate oxide breakdown is the fact that one BD does not necessarily imply circuit failure (Kaczer et al. 2002).

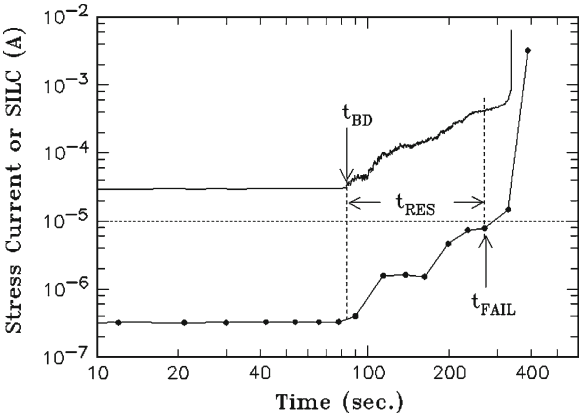


Fig. 2.9 Evolution of the gate current under constant voltage stress (CVS with $V_G = 2.75$ V and T = 140 °C, *top curve*) and the stress-induced leakage (SILC) at 1V (*bottom curve*). The stressed devices are 8 μm^2 nFETs with a $t_{ox} = 1.35$ nm. t_{BD} indicates the first soft breakdown (SBD) event, t_{FAIL} marks the initiation of the hard breakdown (HBD) effect. The residual time (t_{RES}) between soft and hard breakdown depicts the progressive breakdown state. *Source* (Sune et al. 2006)

Bias Temperature Instability

Bias Temperature Instability (BTI) recently gained a lot of attention due to its increasingly adverse impact in nanometer CMOS technologies (Schroder and Babcock 2003). BTI is typically observed as a V_{TH} shift after a bias voltage has been applied to a MOS gate at elevated temperature. For example, when measured over a lifetime of 5 years and under normal operating conditions, V_{TH} shifts of up to 30 mV can be expected for transistors processed in a sub-45 nm technology (Kaczer et al. 2010). BTI-induced degradation of the carrier mobility has also been measured (Schroder and Babcock 2003).

Two different BTI phenomena can be observed: negative BTI (NBTI) and positive BTI (PBTI). NBTI occurs in pMOS transistors when a negative bias voltage is applied. This effect is a significant reliability threat in both older SiO_2 and SiON technologies and is still a problem in newer HKMG technologies (Degraeve et al. 2008). The PBTI effect affects nMOS transistors and results in a similar wearout behavior as NBTI, but has only been observed in HKMG nMOS devices. There, the impact of PBTI on the transistor characteristics can be similar to or even larger than the NBTI effect (Grasser et al. 2010). Currently, there still is no consensus about the microscopical origins of both BTI phenomena. Most authors argue that the NBTI effect results from a combination of hole trapping in oxide defects and generation of interface states at the channel oxide interface (Schroder and Babcock 2003; Kaczer et al. 2008; Grasser and Kaczer 2009). PBTI is believed to come from electron trapping in preexistant oxide traps, combined with a trap generation process (Crupi et al. 2005; Ioannou et al. 2009). Further, initial research on next generation CMOS structures such as multi-gate devices (MuGFETs, FinFETs, etc.) indicates that BTI remains a problem in future CMOS technologies (Groeseneken et al. 2008; Wang et al. 2011; Feijoo et al. 2012).

When time-dependent voltage stress is applied, a peculiar property of the BTI mechanism is revealed: the so-called relaxation or recovery of the degradation immediately after the stress voltage has been reduced (see Fig. 2.10) (Kaczer et al. 2008). This phenomenon greatly complicates the evaluation of BTI, its modeling, and the extrapolation of its impact on circuits. It currently appears that BTI degradation does not fully recover when the stress is removed, hence leaving a permanent residual degradation. BTI degradation can therefore be modeled as a combination of a permanent and a recoverable degradation component (Grasser and Kaczer 2009; Maricau et al. 2011):

$$\Delta V_{TH} \propto \left[\underbrace{\exp\left(\alpha_1 V_{GS}\right) t^{n_P}}_{\text{Permanent Part}} + \underbrace{V_{GS}^{\alpha_2}\left(C_R + n_R \log_{10}(t)\right)}_{\text{Recoverable Part}} \right] \exp\left(-\frac{E_a}{kT}\right) \tag{2.6}$$

where ΔV_{TH} is a function of the transistor gate-oxide electric field (E_{ox}) and the temperature (T). Further, α_1, α_2 are technology-dependent voltage scaling factors, E_a is the activation energy, C_R, n_P and n_R are the time exponents for the permanent

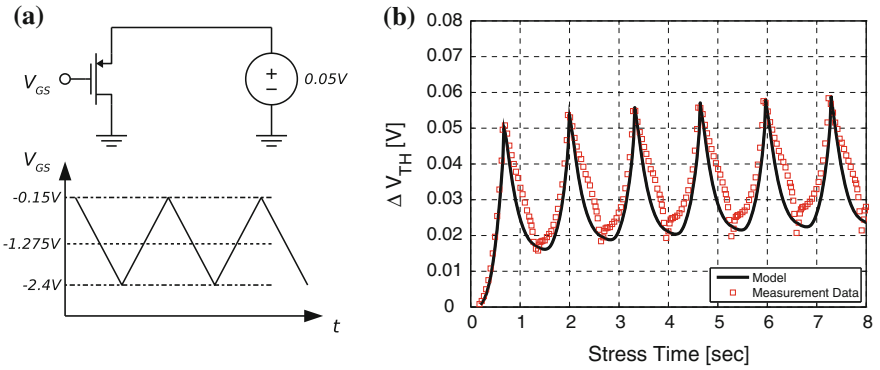


Fig. 2.10 The time-dependent V_{TH} shift of a pMOS transistor subjected to a triangle-shaped stress voltage (see **a**). Part of the NBTI damage is recovered, every time the stress is reduced (see **b**) (Maricau et al. 2011)

and recoverable part and k is the Boltzmann constant. Note how Eq. (2.6) is only valid for a fixed stress voltage, an accurate BTI model for time-varying stress voltages is given in Chap. 3. Also, it is important to note that BTI is shown not to be frequency-dependent (i.e. at least for measurements up to 3 GHz) (Sasse 2008; Ramey et al. 2009). Further, BTI drain bias dependency has also been observed (Schlünder et al. 2005; Luo et al. 2007).

BTI effects in large micrometer-sized transistors are typically considered deterministic (Wang et al. 2007; Maricau et al. 2011). The application of a given stress on matched transistors therefore results in an identical shift of the transistor parameters. Scaling transistors down to nanometer dimensions, however, gradually changed these deterministic effects into stochastically distributed failure mechanisms due to an ever-increasing impact of individual trapping and detrapping events (Kaczer et al. 2010, 2011) (see Fig. 2.11). At device level this results in a time-dependent shift of the transistor parameters ($\Delta V_{TH} = f(t)$) augmented with a time-dependent increase of the standard deviation on these parameters ($\sigma(V_{TH}) = g(t)$). Initially matched transistors, processed in ultra-scaled nanometer CMOS technologies, can therefore cause circuit performance failure resulting from increased time-dependent transistor mismatch (Gielen et al. 2011).

Electromigration

Electromigration (EM) is an aging effect taking place in interconnect wires, contacts and vias in an integrated circuit (Tu 2003). The effect causes material transport by gradual movement of the ions in a conductor due to the momentum transfer between conducting electrons and the diffusing metal atoms. EM is important in applications where high direct current densities are used. Integrated circuits are very prone to this

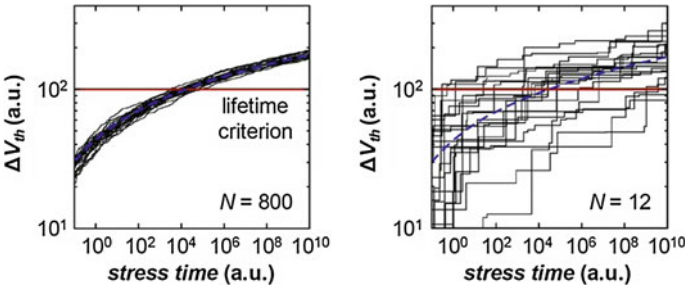


Fig. 2.11 Due to the discrete nature of trapping and detrapping events, the time-dependent BTI V_{TH} shift becomes stochastic for small devices in sub-45 nm CMOS. The lifetime of a large transistor with 800 defects (*left*) is much better defined than the lifetime of a small transistor with only 12 defects (*right*) (Kaczer et al. 2011)

effect, since current densities in excess of 1e5A/cm^2 are being measured (Tu 2003; Lewyn et al. 2009). A typical household extension cord carries only about 1e2A/cm^2 as it is limited by Joule heating rather than electromigration. The electromigration phenomenon is already known for over 100 years, but first became a practical problem in 1966 when the first integrated circuits became commercially available. Figure 2.12 illustrates how the gradual shift of the metal can create a void (open) or a hillock (short) which can potentially cause circuit failure.

In a homogeneous crystalline structure, there is hardly any momentum transfer between the conducting electrons and the metal ions. However, at the grain boundaries, this homogeneity does not exist and the conducting electrons have a large impact on the metal ions. This causes atoms to become separated from the grain boundaries and to be transported in the direction of the current, along the grain boundaries (Jerke and Lienig 2004). The mean time-to-failure (MTTF) of a wire, when subjected to electromigration, can be expressed by Black's law (Black 1969):

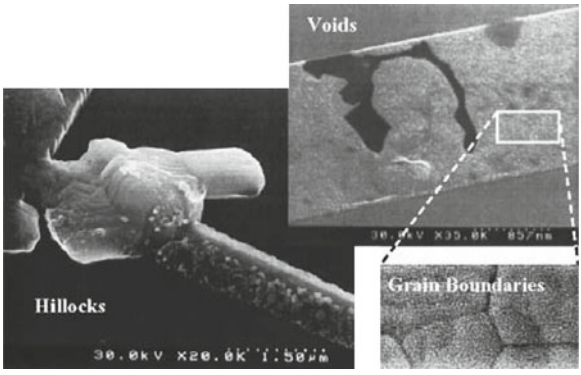


Fig. 2.12 Hillock and void formations in wires due to electromigration (Jerke and Lienig 2004)

$$\mathrm{MTTF} = \frac{A}{J^n} \exp\left(\frac{E_\mathrm{a}}{kT}\right) \tag{2.7}$$

with A a constant dependent on the cross-sectional area of the interconnect wire, J the current density, E_a the activation energy (e.g. 0.7 eV for aluminum), k the Boltzmann constant, T the temperature and n a scaling factor (typically $n = 2$). Note how, besides the current density, the temperature also strongly affects the lifetime of the wire. For an interconnect to remain reliable at high temperatures, the maximum current density must decrease. EM is a very layout dependent phenomenon. The MTTF of a wire does not only depend on the width of the wires, but particular attention must also be paid to vias and contact holes. Since the ampacity of a (tungsten) via is less than of a metal wire, a via is more prone to EM compared to a wire with the same dimensions. Where needed, multiple vias must therefore be used. Also, these vias must be organized such that the current is distributed evenly through all the vias. Additionally, 90-degree corner bends in wires must be avoided, since the current density in such bends is higher than that in oblique angles (Jerke and Lienig 2004; Lienig and Jerke 2005).

In older technologies, aluminum was commonly used as a conducting material for interconnect wires. Aluminum has a good conductivity and a good adherence to the silicon substrate. However, aluminum is very susceptible to electromigration. Research indicated how adding 1–2 % of copper to aluminum increased the resistance to EM about 50 times. This effect is attributed to the fact that copper inhibits the diffusion of atoms along the grain boundaries (Tu 2003). Due to the further scaling of CMOS technologies, a need for a better interconnect conductor than Al(Cu) (having a lower resistance-capacitance delay) was needed. Therefore the industry has turned to full-copper interconnect wires. Copper has a much higher melting point than aluminum and therefore atomic diffusion should be much slower in copper than in aluminum. So, electromigration should be much less in copper interconnects. Surprisingly, the benefit is not as big as expected, and when compared to Al(Cu) wires, copper wires have a lower MTTF. As a solution, tin has been found very effective in retarding electromigration in copper (Tu 2003). However, electromigration still remains a major problem in nanoscale CMOS circuits today (Zhang et al. 2010).

2.4.2 Transient Effects

Transient unreliability effects distort the normal operation of a circuit for a limited time period. Typically the term signal integrity (SI) is used to describe how the quality of a signal in an electronic system changes under these transient effects. Is the signal properly transferred from one subcircuit to the next and what is the quality of the signal at the circuit output? A good-quality signal guarantees high-speed reliable data transfer within a system and between different systems. A signal waveform in an integrated circuit can be distorted by two types of unwanted signals: noise and interference.

Noise

Noise is an unwanted and random perturbation of a signal and results from active or passive devices (e.g. thermal noise, flicker noise, popcorn noise, etc.) within the circuit itself. Noise is bounded by physical limitations and influenced by the fabrication process (technology, device selection, processing quality, etc.). Since device noise determines the minimum detectable signal level, the operation of analog circuits in particular is very prone to these noise sources. Noise is typically modeled as an input-referred noise source, determined from circuit noise analysis and quantified using the noise figure (NF) and signal-to-noise ratio (SNR) parameters. Noise is the ultimate limit to performance in electronic circuits.

Electromagnetic Interference

Electromagnetic interference (EMI) is defined as the influence of unwanted signals generated by a source circuit and picked up by a receptor or victim circuit, affecting its signal performance and quality. The coupling path between the source and the victim circuit can be conductive, capacitive, magnetic or radiative. A coupling path can also consist of two or more of these coupling mechanisms working together. As opposed to noise, an electromagnetic signal has a source or origin external to the (sub)circuit it affects. The source signal can be deterministic (man-made) or random (natural). Examples of natural electromagnetic interference sources are atmospheric noise (e.g. produced by lightning during thunderstorms) and cosmic noise. Man-made interference signals can be functional signals, which are generated during the normal operation of the source circuits, or accidental signals. Examples of man-made interference are on-chip crosstalk and simultaneous switching noise (functional EMI caused by other circuits that are part of the system) and mobile devices, engine ignitions and microwave ovens (accidental EMI caused by unrelated external sources) (Redoute 2009; Loeckx 2010).

1. **On-chip crosstalk** between two circuits or circuit elements (e.g. interconnect wires) is defined as a deviation from the ideal signal waveform propagating in the victim circuit, caused by the influence of signal transitions in the source circuit. Basically, three types of parasitic coupling may result in crosstalk: electric field coupling, magnetic field coupling and common impedance coupling. The latter occurs when multiple current paths share the same conductor. If the source circuit generates noise in the conductor, in the form of alternating current, the voltage over the finite-impedance conductor is modulated. Therefore, the current going to the victim circuit is also modulated and the operation of the victim circuit might be affected. Electric field coupling, on the other hand, results from capacitive coupling between different interconnect nets. In the same way, magnetic coupling can be modeled by coupled inductors (Redoute 2009).
2. **Simultaneous switching noise (SSN)** is a particular case of common impedance crosstalk when subcircuits on the same IC share the same power distribution bus.

This phenomenon is also known as power/ground noise, ground bounce, substrate noise or dI/dt noise when the power/ground signal is connected with a bondwire. Simultaneous switching of multiple digital gates produces large transient current spikes which flow through the power and ground lines of the chip (Redoute 2009). In case of a mixed-signal circuits, SSN is the primary source of substrate noise, where interference generated by the digital circuit influences the operation of a neighboring and potentially sensitive analog circuit (Donnay and Gielen 2003; Stefanou 2011).

3. **Energetic particles** such as alpha and gamma particles result in ionized radiation of the semiconductor material and potentially cause a non-destructive change in the state of CMOS devices. Naturally occurring alpha particles, impinging on the transistors in a circuit, generate electron-hole pairs in several picoseconds. The charges generated in or near the depletion region are separated by the oxide electric field. These particles are particularly dangerous for storage devices or memories as they initiate state changes, resulting in a soft error or single event upset (SEU).
4. **Radiated EMI** is a form of electromagnetic interference where a remote source (e.g. another circuit, a cellphone antenna, a running engine or a microwave oven) becomes an unintentional transmitter of electromagnetic waves that are picked up by the victim circuit. The receiver antenna in the victim circuit can be formed by PCB traces, connection cables or wire loops in the integrated circuit. Once picked up, the irradiated signal can disturb the normal operation of the victim circuit.

Typically, a circuit is subjected to various sources of electromagnetic interference at the same time. The power and frequency spectrum of these interference sources can also vary with the environment, the temperature and the circuit workload (e.g. in case of interconnect crosstalk or substrate noise). It is therefore, at design time, very hard to predict the impact of unwanted interference signals on the operation of the circuit. To guarantee reliable circuit operation, electromagnetic compatibility (EMC) regulations for both emission of (EME) and susceptibility to (EMS) interference signals are used. Each circuit, depending on the field of application, must comply to these rules. The international electrotechnical commission (IEC), for example, is one of the international standards organizations which are addressing the need for standardized IC EMC test methods, such as the IEC 61000 standard (IEC 61000 structure 2012).

2.5 Conclusions

This chapter has briefly reviewed the major unreliability effects affecting the correct operation of circuits integrated in a nanometer CMOS technology. A distinction between spatial and temporal unreliability effects has been made. The former are visible right after fabrication and include random dopant effects, line edge roughness and oxide thickness variations as dominant phenomena in sub-65 nm CMOS.

Temporal unreliability effects are time-dependent and include aging effects (e.g. bias temperature instability and electromigration) and transient effects (e.g. noise and electromagnetic interference). The description of the effects in this chapter, though being far from complete, has summarized the most important aspects of CMOS reliability and additional references have been provided for the interested reader. The remainder of this work focuses on temporal transistor aging effects, although spatial random effects are also included in the circuit reliability simulator proposed in Chap. 5.

Chapter 3
Transistor Aging Compact Modeling

3.1 Introduction

The focus of this work is on simulation and analysis of the impact of transistor aging on ICs integrated in nm CMOS processes. Accurate circuit simulation starts with the availability of good transistor compact models. This chapter therefore discusses the development of a set of models for simulation of the most important aging effects. Most models present in literature focus on understanding the underlying physical effects related to transistor aging, rather than on developing a good compact model. Such a compact model should not only be accurate, but also easy to calibrate and easy to evaluate. Further, it should include all important circuit parameters such as applied stress voltages, transistor dimensions, temperature, etc.

This chapter discusses the development of a transistor compact model for the most important transistor aging effects. First, Sect. 3.2 discusses existing models for hot carrier injection and then proposes a new model optimized for analog circuit simulation. Similarly, Sect. 3.3 first focuses on existing models for bias temperature instability (NBTI). These models, however, are typically too complex for circuit simulation or do not properly include the impact of time-varying stress. Therefore, a new NBTI compact model, solving these problems, is proposed. Next, Sect. 3.4 focuses on models for time-dependent dielectric breakdown. Section 3.5 then proposes an aging-equivalent transistor model to enable simulation of the combined impact of all the different aging effects. To enable a designer to do get a quick estimate of the impact of aging on a circuit, Sect. 3.6 also discusses a first-order aging model for hand calculations. The conclusions of this chapter are given in Sect. 3.7.

E. Maricau and G. Gielen, *Analog IC Reliability in Nanometer CMOS*,
Analog Circuits and Signal Processing, DOI: 10.1007/978-1-4614-6163-0_3,

3.2 Hot Carrier Injection

Transistor damage due to hot carrier injection (HCI) was a major reliability problem in the early eighties (Takeda et al. 1983; Tam et al. 1984; Hu et al. 1985). Later, when supply voltages were scaled down and graded drain junctions were introduced, HCI became a less dominant reliability problem. However, HCI can still pose a problem for circuits processed in high voltage or older CMOS technologies (i.e. >90 nm) (Moens et al. 2010). Even analog circuits processed in more advanced nodes could be at risk, especially when large voltages are applied (e.g. inductor-based oscillators or power amplifiers) (Chouard et al. 2010; Sagong et al. 2011). Therefore, it is still important to correctly estimate the impact of HCI on the behavior of a circuit. This requires an accurate HCI compact model. HCI mainly occurs in nMOS transistors and causes a shift of important transistor parameters such as the threshold voltage and the carrier mobility.

This section first reviews the most important HCI models published in literature. Then, a new compact model, intended for reliability simulation of analog circuits, is proposed. To facilitate calibration and evaluation, the model uses design-related parameters such as transistor voltages and currents. Further, the model supports the evaluation of time-varying stress voltages, which is crucial when analyzing analog circuits. The model is calibrated in a 65 nm CMOS technology. A final section discusses the impact of HCI in sub-45 nm CMOS technologies.

3.2.1 Background

Of all transistor aging effects, HCI is most thoroughly investigated and documented in literature. As a consequence, a large number of compact models have been developed. This section reviews some of the most important and most commonly models used for circuit lifetime evaluation.

The Lucky Electron Model

Most HCI compact models available in literature are based on the 'lucky electron' model (LEM). The concept was first introduced by Shockley in 1961 to explain bulk phenomena (shockley 1961) and later became very popular as the underlying mechanism in a lot of HCI models (Takeda et al. 1983; Hu et al. 1985; Leblebici and Kang 1993; Kufluoglu and Ashraful Alam 2004; Wang et al. 2007). The lateral electric field near the drain region of a transistor is considered as the main reason behind the HCI phenomenon. First, when carriers are emitted out of the source, they are accelerated in the channel. Then, near the drain region, some of those carriers (also called 'lucky' carriers or 'lucky electrons' in case of nFET devices) are injected into the gate oxide, under the influence of the large lateral electric field near the drain.

There, they create interface states and oxide charges resulting in a shift of important transistor parameters such as the threshold voltage. This HCI mode is also referred to as channel hot carrier (CHC) injection and is the most dominant HCI mechanism for longer channel devices that operate in the low V_{GS} regime (Li et al. 2008).

Hu et al. (1985) was one of the first to introduce a HCI model based on the lucky electron concept. Most HCI models published later are based on the same theory but propose a different analytical formulation or are intended to accommodate the model to more advanced CMOS technologies. For example, Kufluoglu and Ashraful Alam (2004) unified the HCI and NBTI effect into one model, based on the generation of interface traps due to Si–H bond dissociation. This theory was used later by Wang et al. (2007) to develop a HCI model for simulation of digital circuits processed in advanced nanometer CMOS technologies. In this semi-empirical model, HCI is expressed in terms of the number of generated interface states ΔN_{IT}:

$$\Delta N_{\mathrm{IT}} = C_1 \left[\frac{I_{\mathrm{DS}}}{W} \exp\left(-\frac{\phi_{\mathrm{IT},e}}{q \lambda_e E_{\mathrm{lat}}} \right) T_{\mathrm{str}} \right]^n \tag{3.1}$$

where W is the transistor channel width, I_{DS} is the drain-source current, E_{lat} is the peak lateral electric field at the drain, T_{str} the stress time, $\phi_{\mathrm{IT},e}$ the critical energy for electrons to create an interface trap ($\phi_{\mathrm{IT},e} \approx 3.7\,\mathrm{eV}$ (Hu et al. 1985)), λ_e is the hot-electron mean-free path ($\lambda_e \approx 6.7\,\mathrm{nm}$ (Leblebici and Kang 1993)) and C_1 and n are process-dependent constants ($C_1 \approx 2$ (Leblebici and Kang 1993) and $n \approx 0.5$ (Wang et al. 2007; Maricau et al. 2008)). In Eq. (3.1), E_{lat} is the most important parameter, but is difficult to obtain accurately with an analytical model. Therefore, especially in earlier models, HCI stress is typically captured as a function of the substrate current I_{sub}:

$$\Delta N_{\mathrm{IT}} = C_2 \left(\frac{I_{\mathrm{sub}}}{W} \right)^{\alpha} T_{\mathrm{str}}^n \tag{3.2}$$

where C_2 and α are technology-related constants. Temperature acceleration is often regarded as a minor effect in most HCI models, but can be modeled with an additional Arrhenius temperature factor:

$$\Delta N_{\mathrm{IT}} = C_3 \left(\frac{I_{\mathrm{sub}}}{W} \right)^{\alpha} \exp\left(\frac{E_{\mathrm{a}}}{kT} \right) T_{\mathrm{str}}^n \tag{3.3}$$

where the activation energy E_{a} is typically around $-0.05\,\mathrm{eV}$ (Lunenborg 1996; Maricau et al. 2008). Eventually, the number of generated interface states, can be related to a shift in transistor parameters such as the threshold voltage V_{TH} and the carrier mobility μ (Sun and Plummer 1980):

$$V_{TH} = V_{TH,0} + \frac{qN_{IT}}{C_{ox}} \tag{3.4}$$

$$\mu = \frac{\mu_0}{1 + \beta \Delta N_{IT}} \tag{3.5}$$

where q is the magnitude of the electric charge of an electron ($q = 1.602\text{e}{-19}\text{C}$), C_{ox} is the oxide capacitance per unit area. $V_{TH,0}$ and μ_0 are the threshold voltage and carrier mobility for an unstressed transistor respectively. β is a process-dependent parameter.

Alternative Models

Although most HCI models are based on the LEM, these models only include one source of hot carriers. However, as discussed in Sect. 2.4, three other HCI mechanisms have also been identified are can potentially result in a time-dependent shift of the transistor behavior. To account for these effects, more advanced HCI models have been published in literature. For example, Li et al. (2008) proposed an improved model that includes the three most important sources of hot carriers: channel hot electrons, substrate hot electrons and drain avalanche hot carriers, depending on the drain current, the substrate current and the drain voltage respectively. Tudor et al. (2011) further improved on Li's model by combining the separate effects into one model that is equally accurate but easier to calibrate.

Further, when scaling to deep-submicron technologies and working at lower voltages, different and more accurate explanations for the HCI effect have been proposed. In 2001, Rauch et al. suggested electron–electron scattering as the dominant reaction behind HCI in nanometer CMOS nFETs. That model was further refined in (Rauch and La Rosa 2005). Guerin et al. (2007) confirmed this energy-driven approach and they showed how, in the high-energy long-channel case, this approach allows to retrieve LEM-like equations. Further, when the energy is lowered, high-energy electrons are mostly generated by electron–electron scattering. Finally, for even lower energies, the multiple vibration excitation mechanism dominates the hot carrier degradation rate. This understanding led to the development of new and more complete models to estimate HCI in advanced CMOS processes (Bravaix et al. 2009).

3.2.2 A HCI Compact Model for Circuit Simulation

In Sect. 3.2.1 an overview of existing HCI models has been given. Over a scope of 30 years, various HCI models have been developed. Starting from models based on the 'lucky electron' concept in the eighties towards more advanced models based on an energy-driven approach early in the twenty-first century. However, most of these models are intended to better understand the physical mechanisms behind the HCI

phenomenon, rather than to provide a compact model for circuit simulation. As a consequence, the industry standard for HCI effects analysis remained based on the I_{sub}-based LEM, as proposed by Hu et al. (1985) (also see Eq. (3.2)). This model is however no longer valid for advanced nanometer CMOS processes, since I_{sub} is then dominated by other leakage components such as gate leakage, junction current and gate-induced drain leakage (Wang et al. 2007). Therefore, Wang et al. (2007) proposed a HCI compact model based on the LEM, but optimized for digital circuit reliability simulation in nanometer CMOS. In this section, a alternative HCI compact model for analog circuit simulation in sub-250 nm CMOS is discussed.

A DC Model

The model is developed based on the reaction-diffusion (RD) approach proposed by Kufluoglu and Ashraful Alam (2004) and is also based on the LEM. As will be discussed in Sect. 3.3.2, the RD model has a number of drawbacks and is therefore not well suited to develop a BTI compact model. One of these drawbacks is the absence of support for oxide trapping in the RD model. Oxide trapping, however, is related to the typical BTI recovery effect upon stress removal. Nevertheless, the RD model can still be used to model the hot carrier effect since traps are only generated near the drain end of the transistor and therefore recovery effect is negligible. The RD model consists of a set of two differential equations describing the generation of hydrogen particles near the channel/oxide interface and their diffusion towards the gate contact. These equations can be solved numerically when analyzing only one transistor, but such an approach is not appropriate when analyzing an entire circuit. The computational effort would indeed be too large. Starting from a freshly fabricated transistor subjected to a constant voltage stress, the RD model can be simplified to (Kufluoglu and Ashraful Alam 2004):

$$N_{\mathrm{H}}(0)N_{\mathrm{IT}} \approx \frac{k_{\mathrm{F}}}{k_{\mathrm{R}}}N_0 \tag{3.6}$$

where N_{IT} represents the number of interface traps (i.e. broken Si–H bonds), N_0 is the initial number of unbroken Si–H bonds and k_{F} is the oxide-field-dependent forward dissociation rate constant. The broken Si-bonds act as a donor trap and contribute to the shift in the threshold voltage. $N_H(0)$ is the hydrogen concentration at the interface ($x = 0$) and k_{R} is the annealing rate constant. Hot electrons in the transistor channel induce the breaking of Si–H bonds. As there is no common agreement on how the H-atoms diffuse into the oxide, an artificial H_x diffusion particle is assumed. As such this derivation extends the work described in Kufluoglu and Ashraful Alam (2004) and additionally, it includes explicit dependence on the length of the transistor. The hydrogen reaction at the channel/oxide-interface can be represented by the following equilibrium equation:

$$N_{H_x} = k_{\mathrm{H}}N_{\mathrm{H}}^{n_x} \tag{3.7}$$

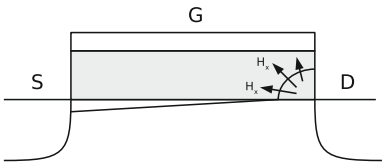


Fig. 3.1 Hydrogen particles H_x result from broken Si–H bonds due to HCI. The hydrogen particles diffuse equally in all directions into the gate oxide

with k_H a reaction constant and n_x the number of hydrogen atoms per hydrogen particle. N_{H_x} is the concentration of hydrogen particles per volume. The number of interface traps can also be calculated by integrating the number of broken Si–H bonds. The hydrogen particles diffuse from the drain, where they are created, into the gate oxide. The diffusion is assumed to be equal in all directions (see Fig. 3.1). The position of the diffusion front is a function of time and can be described as $\sqrt{D_{\mathrm{H}_x} t}$ (Kufluoglu and Ashraful Alam 2004). Every H_x particle consists of n_x H-atoms, therefore the average number of interface traps per unit area can be calculated as:

$$\begin{aligned} N_\mathrm{IT} &= \frac{\pi W}{2A_\mathrm{tot}} n_\mathrm{x} \int_0^{\sqrt{D_{\mathrm{H}_x} t}} \left(N_{\mathrm{H}_x}(0) \left[r - \frac{r^2}{\sqrt{D_{\mathrm{H}_x} t}} \right] \right) dr \\ &= N_{\mathrm{H}_x}(0) \frac{\pi n_x}{12L} D_{\mathrm{H}_x} t \end{aligned} \tag{3.8}$$

where W and L represent the width and length of the transistor and A_tot is the total area under the transistor gate. Combining (3.6), (3.7) and (3.8) eventually results in:

$$N_\mathrm{IT} = \left(\frac{k_\mathrm{F} N_0}{k_\mathrm{R}} \right)^{\frac{n_x}{1+n_x}} \left(\frac{n_\mathrm{x} \pi k_\mathrm{H}}{12L} D_{\mathrm{H}_x} \right)^{\frac{1}{1+n_x}} t^{\frac{1}{1+n_x}} \tag{3.9}$$

The number of interface traps N_IT is directly related to a shift in transistor performance parameters such as the threshold voltage (also see Sect. 3.5), therefore:

$$\Delta V_\mathrm{TH} = \underbrace{C_\mathrm{HCI} \left(\frac{k_\mathrm{F} N_0}{k_\mathrm{R}} \right)^{\frac{n_x}{1+n_x}} \left(\frac{n_x \pi k_\mathrm{H}}{12L} D_{\mathrm{H}_x} \right)^{\frac{1}{1+n_x}}}_{A_\mathrm{HCI}} t^{\frac{1}{1+n_x}} \tag{3.10}$$

with C_HCI a technology-dependent parameter. To use this model in a circuit simulator, physics-related model parameters in Eq. (3.10) have to be related to observable transistor parameters such as voltages, currents and transistor dimensions:

1. The forward reaction constant, k_F, is a function of the transistor temperature and the biasing voltages of the transistor (Alam and Mahapatra 2005; Wang et al. 2007).

$$k_{\mathrm{F}} \propto C_{\mathrm{ox}}(V_{\mathrm{GS}} - V_{\mathrm{TH}}) \cdot \exp\left(\frac{E_{\mathrm{ox}}}{E_0}\right) \cdot \exp\left(\frac{-\phi_{\mathrm{IT}}}{q\lambda E_{\mathrm{lat}}}\right) \cdot \exp\left(\frac{-E_{k_{\mathrm{F}}}}{kT}\right) \quad (3.11)$$

The first term in (3.11) represents the number of carriers in the channel, the second term expresses the dependence of k_F on the gate oxide field E_{ox} with E_0 a technology-dependent constant. E_{ox} is a function of V_{GS}. The third term represents the probability for an electron to gain an energy ϕ_{IT} or more in an electric field E_{lat}. ϕ_{IT} is the minimum impact ionization energy in electronvolts. λ is the mean free path of an electron travelling through the channel. E_{lat} is the maximum horizontal electric field in which the electrons accelerate and depends both on V_{DS} and V_{GS}. Finally, the last term in Eq. (3.11) is an Arrhenius temperature dependency with activation energy E_{k_F}. A higher temperature increases the silicon lattice vibration, which decreases the free path length of an accelerating electron in the pinch-off region and thus reduces the number of hot carriers. This results in a negative activation energy E_{k_F}.

2. The reverse reaction constant, k_R, and the diffusion constant, D_{H_x}, only depend on the temperature via an Arrhenius factor:

$$k_{\mathrm{R}} \propto \exp\left(\frac{-E_{k_{\mathrm{R}}}}{kT}\right) \quad (3.12)$$

$$D_{\mathrm{H}_x} \propto \exp\left(\frac{-E_{D_{\mathrm{H}_x}}}{kT}\right) \quad (3.13)$$

The corresponding activation energies, $E_{k_{\mathrm{R}}}$ and $E_{D_{\mathrm{H}_x}}$ are, in contradiction to $E_{k_{\mathrm{F}}}$, positive instead of negative.

Combining all constant terms in (3.10) and replacing k_{F}, k_{R}, and D_{H_x} by stress and device parameter dependent terms results in a CHC compact model as presented in Table 3.1. E_0, n_x, E_{a} and A are to be determined for every process through measurements. From Eq. (3.10) and Table 3.1 one can see how the analytic expression proposed here has a time-dependent behavior of the form $A_{\mathrm{HCI}}t^{n_{\mathrm{HCI}}}$ with A_{HCI} depending on the stress conditions and $n_{\mathrm{HCI}} \approx 0.5$. This is in good agreement with other models proposed in literature (Hu et al. 1985; Kufluoglu and Ashraful Alam 2004). Note that for large V_{TH} shifts, the length of the pinch-off region (parameter l in the model) decreases (Wong et al. 1997). This results in an increasing number of channel interface traps, decreasing the carrier mobility. Further, parameter l also changes for varying transistor lengths and drain voltages (Wong et al. 1997).

The HCI model discussed above was verified experimentally on an IMEC 65 nm process. A transition method using the subthreshold to strong inversion transition region is adopted to extract the V_{TH} (Garcia sanchez et al. 2000). Each device is stressed over a period of 9000 s and the threshold voltage is extracted during measurement phases that increase geometrically over time. The model peters, extracted from the measurement data, are also given in Table 3.1. Figure 3.2 compares the model with measurement results, for various combinations of the gate and drain

Table 3.1 DC channel HCI compact model and parameter values extracted for a 65 nm CMOS process

ΔV_{TH}	$A_{\mathrm{HCI}} t^{\frac{1}{1+n_x}}$		
A_{HCI}	$C_{\mathrm{HCI}} \left[(V_{\mathrm{GS}} - V_{\mathrm{TH}}) K_v\right]^{\frac{n_x}{1+n_x}} \left(\frac{n_x}{L}\right)^{\frac{1}{1+n_x}}$		
K_v	$\exp\left(\frac{E_{\mathrm{ox}}}{E_0}\right) \exp\left(\frac{-\phi_{\mathrm{IT}}}{q\lambda E_{\mathrm{lat}}}\right) \exp\left(-\frac{E_{\mathrm{a}}}{kT}\right)$		
E_{ox}	$\frac{(V_{\mathrm{GS}} - V_{\mathrm{TH}})}{t_{\mathrm{ox}}}$		
E_{lat}	$\frac{(V_{\mathrm{DS}} - V_{\mathrm{DSAT}})}{l}$		
V_{DSAT}	$\frac{E_{\mathrm{sat}} L (V_{\mathrm{GS}} - V_{\mathrm{TH}})}{E_{\mathrm{sat}} L + (V_{\mathrm{GS}} - V_{\mathrm{TH}})}$		
E_{sat}	$\frac{2v_{\mathrm{sat}}}{\mu_{\mathrm{eff}}}$		
μ_{eff}	$\frac{\mu_0}{1 + \theta(V_{\mathrm{GS}} - V_{\mathrm{TH}})}$		
C_{HCI}	1.5e–5	n_x	1.21
E_0 (V/m)	0.71e8	E_{a} (eV)	−0.06
t_{ox} (m)	2.2e–9	ϕ_{IT} (eV)	3.7
λ (m)	7.8e–9	l (m)	45e–9
v_{sat} (m/s)	1e5	μ_0 (cm^2/Vs)	235
θ (V^{-1})	0.95	k (J/K)	1.38e–23
$V_{\mathrm{TH,0}}$ (V)	0.2	q (C)	1.602e–19

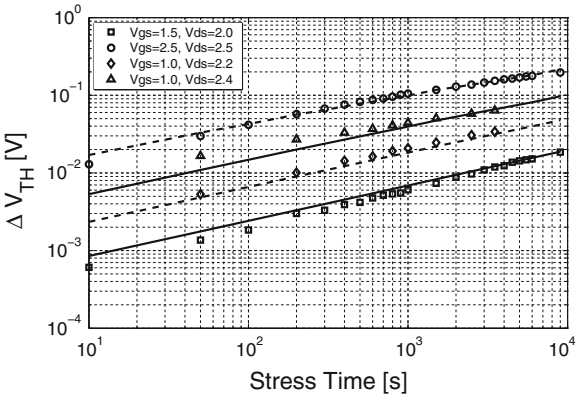


Fig. 3.2 Validation of the DC HCI model for various combinations of the gate and drain voltages. The model (*lines*) fits the measurement data (*markers*) very well

voltages. The model fits the measurement data very well. Figure 3.3 depicts the measurement results for three measurements on different devices but under the same stress conditions ($V_{\mathrm{GS}} = 1.5\,\mathrm{V}$ and $V_{\mathrm{DS}} = 2.0\,\mathrm{V}$). Variation on the measurement data is small, compared to the HCI induced ΔV_{TH}, and is assumed to mainly result from process-induced fluctuations in the gate dimensions.

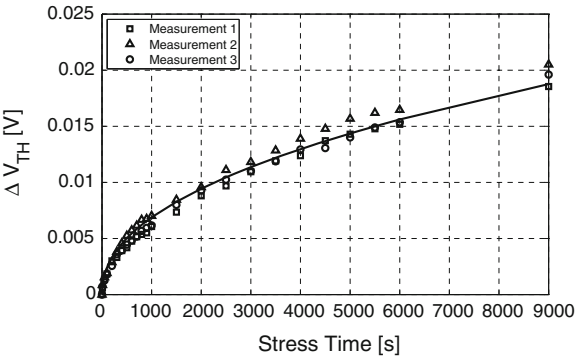


Fig. 3.3 A comparison between three replica measurements on different but identically sized transistors ($V_{GS} = 1.5\,V$ and $V_{DS} = 2.0\,V$). Variation on the measurement data is small, compared to the HCI induced ΔV_{TH}, and is assumed to mainly result from process-induced fluctuations in the gate dimensions

An AC Model

The model, proposed in the previous section and depicted in Table 3.1, is only valid for DC stress voltages (i.e. when a fixed stress is applied). Transistors integrated in an actual circuit, however, are typically subjected to time-varying voltages. To also include time-varying stress, one can derive a simple extension to the DC model. First, a transistor is stressed with a fixed stress voltage during a period t_1, resulting in a threshold voltage shift equal to:

$$\Delta V_{TH,1} = A_{HCI,1} t_1^{n_{HCI}} \tag{3.14}$$

$A_{HCI,1}$ is a function of the applied stress voltage and the temperature during t_1. An expression for $A_{HCI,1}$ is given in Table 3.1. Then, the stress voltage is changed to $A_{HCI,2}$ during a period $T_2 = t_2 - t_1$. Now, one can find an equivalent stress time $t_{1,eq}$ for $A_{HCI,2}$, which results in a HCI degradation equal to the damage done by $A_{HCI,1}$ during t_1:

$$\begin{aligned} A_{HCI,1} t_1^{n_{HCI}} &= A_{HCI,2} t_{1,eq}^{n_{HCI}} \\ t_{1,eq} &= \left(\frac{A_{HCI,2}}{A_{HCI,1}} t_{1,eq}^{n_{HCI}} \right)^{1/n_{HCI}} \end{aligned} \tag{3.15}$$

Now, the total damage after time t_2 can be calculated as:

$$\Delta V_{TH,2} = A_{HCI,2} \left(t_{1,eq} + T_2 \right)^{n_{HCI}} \tag{3.16}$$

Note how $(t_{1,eq} + T_2) \neq t_2$, if $A_{HCI,1} \neq A_{HCI,2}$. Indeed, an equivalent t_1 is calculated assuming that the transistor was stressed with $A_{HCI,2}$ for the entire time. Further,

this model is only valid because HCI damage can be considered to be permanent (Parthasarathy 2006). Equation (3.16) can also be recast to:

$$\Delta V_{\mathrm{TH},2} = \left(\Delta V_{\mathrm{TH},1}^{1/n_{\mathrm{HCI}}} + A_2^{1/n_{\mathrm{HCI}}}(T_2)\right)^{n_{\mathrm{HCI}}} \tag{3.17}$$

or, for a continuous time-varying stress signal:

$$\Delta V_{\mathrm{TH}}(t) = \left(\int_0^t A_{\mathrm{HCI}}(t)dt\right)^{n_{\mathrm{HCI}}} \tag{3.18}$$

Combining (3.18) and the static model described in Table 3.1, one can evaluate the impact of HCI due time-varying stress signals on the behavior of a transistor processed in a nanometer CMOS process.

3.2.3 HCI in Sub-45 nm CMOS

In sub-45 nm nodes HCI became a renewed problem as the supply voltage scaling is slowing down because of the non-scalability of the subthreshold slope. At the same time, the gate length L is continuing to scale to smaller dimensions (Bravaix et al. 2009). This induces an increase in the lateral electric field E_{lat} and the vertical oxide electric field E_{ox}. This is also clearly shown in Fig. 3.4.

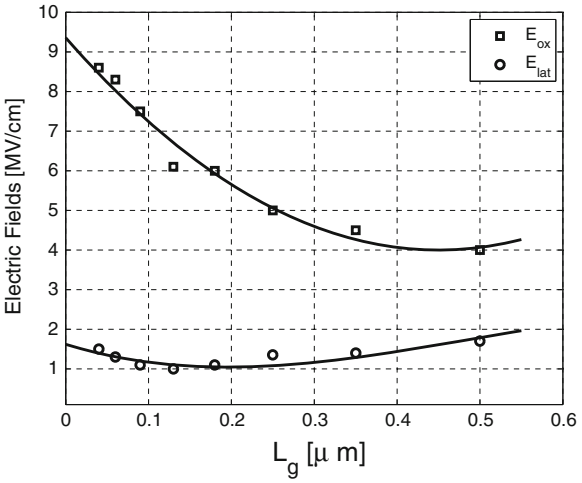


Fig. 3.4 Peak lateral and oxide electric field in commercial nanometer CMOS nodes (Bravaix et al. 2009). Both electric fields increase in the latest CMOS technologies, resulting in a renewed hot carrier reliability concern for these nodes

Further, Magnone et al. (2011) demonstrated a time-dependent increase in the device variability, through measurements performed on a large population of devices (≈1000) processed in a 45 and 65 nm CMOS process. For small devices in these ultra-scaled technologies the number of HCI induced charge traps is very small and does no longer average out. The degradation of a single device is then no longer described as only a change of the average behavior, but also as a change of the standard deviation:

$$\begin{aligned} \mu(\Delta V_{\mathrm{TH}}) &= A_{\mathrm{HCI}} t^{n_{\mathrm{HCI}}} \\ \sigma(V_{\mathrm{TH}}) &= \sqrt{K_{\mathrm{HCI}} \frac{q\mu(\Delta V_{\mathrm{TH}})}{2C_{\mathrm{ox}} WL}} \end{aligned} \tag{3.19}$$

with W and L the dimensions of the transistor. C_{ox} is the oxide capacitance and K_{HCI} a process-dependent constant ($K_{\mathrm{HCI}} \approx 5.6$ for a 45 nm CMOS technology (Magnone et al. 2011)). A compact model to calculate $\mu(\Delta V_{\mathrm{TH}})$ is given in Sect. 3.5.

3.3 Bias Temperature Instability

Negative Bias Temperature Instability (NBTI) is considered to be one of the most critical transistor reliability threats in sub-90 nm CMOS technologies (Degraeve et al. 2008; Lewyn et al. 2009; International technology 2011). Estimating the impact of NBTI wearout effects on the performance of circuits is therefore essential. NBTI, being a temperature-activated effect observed in pFETs, manifests itself as a gradual shift of transistor parameters (e.g. the threshold voltage) when a voltage stress is applied to the transistor gate.

First, this section reviews the most important physical models, trying to explain the NBTI phenomenon. Then, a compact model including all important NBTI dependencies and intended for analog circuit simulation is proposed. This model is calibrated and validated with measurements in a 65 nm CMOS technology. A final section then discusses the impact of PBTI and stochastic BTI effects, both important phenomena in sub-45 nm CMOS technologies.

3.3.1 Background

NBTI was first discovered in the 1960s and over the following 5 decades the effect became one of the most discussed transistor aging effects. Engineers and scientists mainly focused on two important issues: how to properly measure the effect and how to explain the microscopic behavior. Over time, various measurement techniques and associated explanations of the underlying physical behavior have been proposed. In 1977, a first paper suggested a hydrogen-diffusion controlled interface state creation

mechanism. This was called the reaction-diffusion mechanism and was long assumed to be essentially correct, although hydrogen diffusion was suggested to be dispersive later on (see Section). In 2006, however, Huard et al. suggested that NBTI consists of a more or less permanent interface state generation mechanism combined with a recoverable component related to elastic hole trapping. Most scientists first regarded the latter as an experimental nuisance, but then the theory gained more and more momentum.

The Reaction-Diffusion Model

The reaction-diffusion (RD) model is one of the most popular NBTI models and describes the phenomenon as a thermally activated reaction of holes with Si–H bonds at the substrate/dielectric interface of a pMOSFET. The resulting Si dangling bonds give rise to interface states and the free hydrogen particles diffuse away from the interface into the gate dielectric. The RD mechanism was first proposed in 1977 by Jeppson and Svensson (1977). The well-known RD model for NBTI in nanometer CMOS technologies was developed by Alam (2003). Other scientists supported this theory (Schroder and Babcock 2003; Chakravarthi 2004). Originally, this model was used to explain the NBTI gate oxide field and temperature dependency. Later, the model was updated (Alam and Mahapatra 2005; Alam et al. 2007) to also include post-2003 observations, such as the NBTI saturation effect for long stress times and the AC frequency independence.

According to the RD model, NBTI arises due to hole-assisted breaking of Si–H and Si–O bonds at the Si/SiO_2 interface. This electrochemical reaction is field and temperature dependent and the rate of trap generation (i.e. reaction) can be described with the following differential equation:

$$\frac{dN_{IT}}{dt} = k_F(N_0 - N_{IT}) - k_R N_H(0) N_{IT} \tag{3.20}$$

where N_{IT} represents the number of interface traps (i.e. broken Si–H bonds) at any given instant, N_0 is the initial number of unbroken Si–H bonds and k_F is the oxide-field dependent forward dissociation rate constant. The broken Si-bonds act as a donor trap and contribute to the shift in the threshold voltage. The second term in Eq. (3.20) describes how the released hydrogen atoms can also anneal the broken bonds. $N_H(0)$ is the hydrogen concentration at the interface (x = 0) and k_R is the annealing rate constant. Instead of annealing broken bonds, hydrogen atoms can also diffuse away from the interface and into the oxide. This diffusion process is described by:

$$\frac{dN_H}{dt} = D_H \frac{d^2 N_H}{dx^2} \tag{3.21}$$

with N_H the total hydrogen concentration in the oxide and D_H the diffusion constant. In order to find a closed-form analytical expression for the NBTI phenomenon, one

typically assumes that the initial trap generation rate is slow compared to the fluxes on the right hand side of Eq. (3.20). Therefore:

$$N_{\mathrm{IT}} \ll N_0 \sim 5 \times 10^{12}\,\mathrm{cm}^{-2} \tag{3.22}$$

$$\frac{dN_{\mathrm{IT}}}{dt} \approx 0 \tag{3.23}$$

Hence, Eq. (3.20) can be recast to:

$$N_H(0) N_{\mathrm{IT}} \approx \frac{k_{\mathrm{F}}}{k_{\mathrm{R}}} N_0 \tag{3.24}$$

Further, after an initial startup of the reaction-diffusion process, hydrogen diffusion controls the trap generation process. In this regime, the diffusion front is located at (also see Fig. 3.5):

$$x_{\mathrm{DF}}(t) = \sqrt{D_{\mathrm{H}} t} \tag{3.25}$$

Also, since the number of generated interface traps is equal to the number of diffused hydrogen atoms (also see Fig. 3.5):

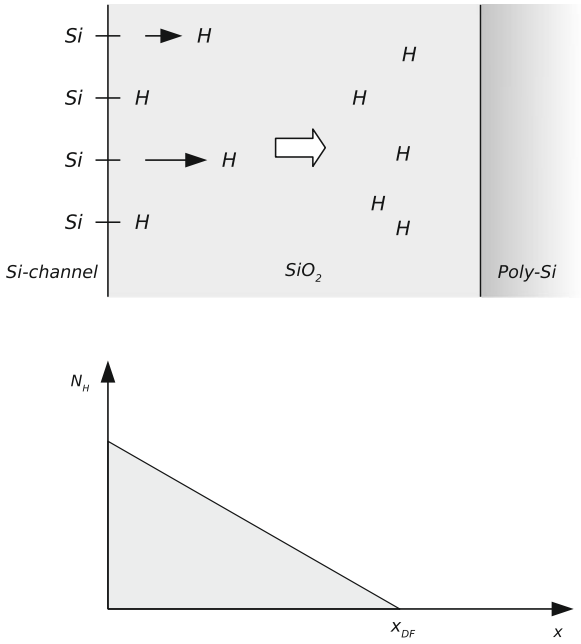


Fig. 3.5 The RD model assumes that NBTI arises due to hole-assisted breaking of Si–H and Si–O bonds at the Si/SiO_2 interface (Alam and Mahapatra 2005)

$$N_{\mathrm{IT}} = \int_0^{\sqrt{D_{\mathrm{H}}t}} N_{\mathrm{H}}(x,t)dx = \frac{1}{2} N_{\mathrm{H}}(0)\sqrt{D_{\mathrm{H}}t} \tag{3.26}$$

Finally, inserting Eq. (3.26) into Eq. (3.24) results in:

$$N_{\mathrm{IT}} = \left(\frac{k_{\mathrm{F}}}{k_{\mathrm{R}}}\frac{N_0}{2}\right)^{1/2} (D_{\mathrm{H}}t)^{1/4} \tag{3.27}$$

The threshold voltage shift, resulting from the increased number of interface traps, can be calculated as:

$$\Delta V_{\mathrm{TH}} = \frac{qN_{\mathrm{IT}}}{C_{\mathrm{ox}}} \tag{3.28}$$

with q the elementary charge of an electron and C_{ox} the gate oxide capacitance. Despite the fact that this model is widely used in academia and in industry, it still has a number of deficiencies:

- The exact value of the time exponent depends on the type of diffusion species. In the original RD model the diffusing species were assumed to be individual hydrogen atoms, resulting in a power-law index of 0.25. Measurements, however, revealed time exponents ranging from 0.25 to 0.11, depending on the applied measurement technique (Grasser et al. 2008b). More recent models therefore often assume that two hydrogen atoms always combine to a H_2 molecule. This results in a time exponent of 0.17 and therefore corresponds better to the measurement data (Wang et al. 2007). Furthermore, to explain field-dependent relaxation, charged particles have been considered. These H^+ species drift through the oxide under the presence of the gate electric field but result in a time exponent of 0.5, which is again not consistent with measurements (Alam et al. 2007).
- The conventional RD model predicts universal recovery. This means that, when the threshold voltage shift ΔV_{TH} is normalized to its value at the end of the stress, it depends only on the ratio of the stress and relaxation times. In particular, neither the forward nor backward reaction rates, nor the diffusion coefficient have any influence on the recovery. This result is in big contrast with experimental facts and evidence has piled up showing that NBTI recovery cannot be the diffusion-limited process suggested by the RD model (Grasser et al. 2011).
- The original RD model, as described here, assumes that interface trap generation is the primary cause of NBTI-related transistor parameter shifts. However, research has shown that hole trapping in preexisting gate-oxide traps cannot be neglected. Moreover, these traps are shown to be responsible for the fast trapping and detrapping of charges, resulting in the typical NBTI recovery behavior under AC stress (Kaczer et al. 2008; Grasser et al. 2009). The failure-time prediction error of analog circuits, typically subjected to time-varying stress, is therefore potentially very large. An updated version of the RD model, including hole trapping, was proposed in (Mahapatra et al. 2011), but a closed-form mathematical expression is still lacking.

- Most model parameters in Eq. (3.27) are not directly linked to common transistor design parameters such as stress voltage and temperature. Indeed, the model's primary purpose is to *explain* the physical NBTI mechanism, rather than to serve as an analytical compact model for reliability circuit simulation. However, Eqs. (3.27) and (3.28) do correspond to a well-known and abundantly used first-order compact model for NBTI (Parthasarathy 2006). This empirical formula is based on measurements and expresses the threshold voltage shift ΔV_{TH} as a function of temperature T, stress voltage V_{GS} and stress time t:

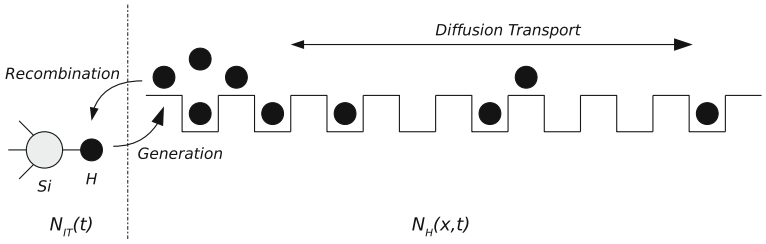
$$\Delta V_{\mathrm{TH}} = A \exp\left(-\frac{E_{\mathrm{a}}}{kT}\right) \exp\left(\gamma V_{\mathrm{GS}}\right) t^{n_{\mathrm{NBTI}}} \tag{3.29}$$

  where A is a technology-dependent constant, E_{a} is the activation energy (typically around 100e–3 eV), k the Boltzmann constant and n_{NBTI} the NBTI time exponent. Model parameters for a 90 nm technology can be found in (Parthasarathy 2006). Unfortunately, Eq. (3.29) is only valid for DC stress voltages. Wang et al. proposed a more advanced analytical NBTI model with parameters extracted for a 65 nm CMOS technology (Wang et al. 2007). This model is based on the RD model and uses empirical formulas to include the relationship between the model parameters and various design parameters such as the stress voltage. The model presented in Wang et al. (2007) is however only suitable for reliability simulation of digital circuits, since the NBTI recovery property is only modeled for square-shaped (i.e. on-off) stress voltages.

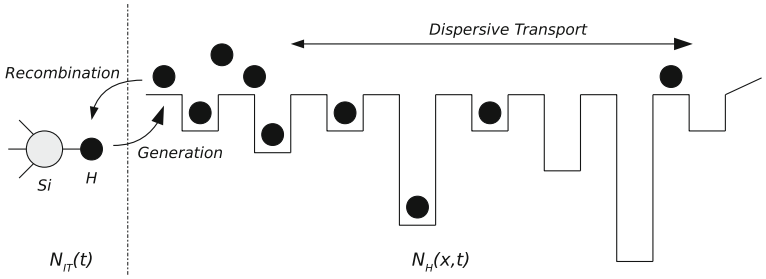
The Reaction-Dispersive-Diffusion Model

Measurements show, even for long stress times of more than 1e3 s, an immediate recovery of more than 60 % of the NBTI damage in the first second after the stress has been removed. The RD model, however, does not predict this fast recovery effect. Since the latter assumes diffusion-limited degradation and relaxation, this would imply a backward diffusion rate that is orders of magnitude faster than the forward diffusion rate which is against the nature of a diffusion process. Drift of charged hydrogen species, added to the RD model later, can also not explain the fast relaxation rate. Indeed, experiments indicate how recovery is independent of the hydrogen passivation degree of the interface (Huard et al. 2010; Aichinger 2010). This is inconsistent with the idea that hydrogen back-diffusion controls recovery, as in that case the hydrogen available at the interface would have an impact (Grasser et al. 2008a).

The reaction-dispersive-diffusion (RDD) model solves this problem by using a trap-controlled movement of hydrogen through the oxide, instead of using a diffusion equation as is done in the RD model (also see Eq. (3.21)) (Kaczer et al. 2005; Zafar 2005; Grasser et al. 2008a). The RDD model can also be seen as an extension of the RD model. It implies that the hydrogen species are separated into two distinct contri-



(a) Reaction-Diffusion Model



(b) Reaction-Dispersive-Diffusion Model

Fig. 3.6 Hydrogen passivated interface states of density N_{IT} are depassivated under negative bias stress. Figure 3.6a: According to the RD model, neutral hydrogen, with concentration $N_{\mathrm{H}}(x,t)$, diffuses into the gate oxide, leaving behind positive interface states. Hydrogen diffusion proceeds via shallow hopping sites in the gate oxide, represented by a regular network of potential wells. Figure 3.6b: The RDD model assumes dispersive transport which is trap-controlled. Diffusing particles spend most of their lifetime in the deep states, which thus control the transport properties (Grasser et al. 2006; Grasser et al. 2008a)

butions: free particles and particles residing on various trap levels. In contradiction to the free particles, the trapped particles do not contribute to the hydrogen transport. Dispersive transport is often described using multiple trapping (MT) models (Grasser et al. 2006). In this model, the total hydrogen concentration N_{H} consists of free (conducting) hydrogen N_{c} and trapped hydrogen N_{t}:

$$N_{\mathrm{H}}(x,t) = N_{\mathrm{c}}(x,t) + N_{\mathrm{t}}(x,t) \tag{3.30}$$

$$= N_{\mathrm{c}}(x,t) + \int \rho(x, E_{\mathrm{t}}, t)\, dE_{\mathrm{t}} \tag{3.31}$$

with E_{t} the trap level energy and $\rho(E_{\mathrm{t}})$ the trapped hydrogen density (expressed in $\mathrm{cm}^{-3}\,\mathrm{eV}^{-1}$) at trap level E_{t}. The continuity equation for the total concentration of the hydrogen species in the oxide, which in the original RD model is described by Eq. (3.21), is now written as (Grasser et al. 2006):

$$\frac{dN_{\mathrm{H}}(x,t)}{dt} = D_{\mathrm{H}} \bigtriangledown^2 N_{\mathrm{c}}(x,t) \tag{3.32}$$

At each trap level, a balance equation accounts for the newly trapped particles versus the released ones. The release rate is proportional to the trapped charge on that level (Grasser et al. 2006):

$$\frac{d\rho(E_{\mathrm{t}})}{dt} = c(E_{\mathrm{t}})N_{\mathrm{c}}\left[g(E_{\mathrm{t}}) - \rho(E_{\mathrm{t}})\right] - r(E_{\mathrm{t}})\rho(E_{\mathrm{t}}) \tag{3.33}$$

where $c(E_{\mathrm{t}})$ and $r(E_{\mathrm{t}})$ are the energy-dependent capture and release rates, respectively. $g(E_{\mathrm{t}})$ represents the trap density of states. When one assumes that only free hydrogen can repassivate dangling bonds, a solution for the MT model can be found. Commonly, the trap density of states is assumed to be exponentially distributed (Grasser 2006):

$$g(E_{\mathrm{t}}) = \frac{N_{\mathrm{t}}}{E_0}\exp\left(-\frac{E_{\mathrm{c}} - E_{\mathrm{t}}}{E_0}\right) \tag{3.34}$$

with E_{c} the hydrogen conduction band edge (which is typically assumed to be zero). Parameter E_0 can be written as:

$$E_0 = \frac{kT}{\alpha} \tag{3.35}$$

with α the dispersion coefficient, k the Boltzmann constant and T the temperature. The hydrogen transport will only be dispersive if α is smaller than one, that is for sufficiently 'deep' trap distributions. From Eqs. (3.33), (3.34) and (3.35), and using the same kinetic equation to describe the interface trap generation rate as for the RD model (see Eq. (3.20)), the number of interface traps N_{IT} can then be calculated (Grasser et al. 2008a). Assuming neutral hydrogen particles, one obtains:

$$N_{\mathrm{IT}}(t) = \left[a\left(\frac{k_{\mathrm{F}}N_0}{k_{\mathrm{R}}}\right)^a\right]^{\frac{1}{1+a}}\left(\frac{D_{\mathrm{H}}}{v_0}\frac{N_{\mathrm{c}}}{N_{\mathrm{t}}}\right)^{\frac{1}{2+2a}}(v_0 t)^{\frac{\alpha}{2+2a}} \tag{3.36}$$

with v_0 the attempt-to-jump frequency. When assuming atomic hydrogen particles, parameter $a = 1$ and the time exponent $n = \alpha/4$. When H_2 is assumed, $a = 2$ and $n = \alpha/6$. In both cases, the time exponent n is increasing for an increasing dispersion parameter α. Since α is one for diffusive transport and zero in the extreme dispersive case, a time exponent smaller than the RD exponents of 1/2, 1/4 and 1/6 can be obtained. Also, for increasing trapped hydrogen concentration N_{t} the total amount of degradation decreases. More details on the RDD model, including alternatives for the MT model to describe the dispersive hydrogen transport and different solutions for various combinations of boundary conditions and initial conditions, can be found in (Grasser et al. 2008a). The RDD model provides a more accurate description of the NBTI phenomenon, but is still not suited for accurate circuit reliability simulation because of the following reasons:

- Similar to the RD model, the coefficients in the RDD model are not directly linked to device parameters such as stress voltages, device sizes and changes in electrical transistor parameters. Further, although this model provides a better description of how interface states are re-passivated, Eq. (3.36) cannot be used as such to calculate transistor degradation when time-dependent stress is applied and solving the differential equations describing the RDD process (see Eqs. (3.20), (3.33) and (3.34)) everytime the stress voltage changes would be computationally too intensive.
- Oxides in modern CMOS processes are too thin to explain a hydrogen diffusion front, as suggested by the RD model. This is explained by assuming that hydrogen mainly diffuses in the polysilicon gate contact (Krishnan et al. 2005). Such an extension is clearly pointless for dispersive hole transport.
- Although interface state generation is a universally acknowledged feature of NBTI, positive charge generation in the oxide bulk has also been reported. This positive charge is related to trapped holes in preexisting traps or in traps generated by hydrogen species (Parthasarathy 2006; Huard et al. 2007; Ielmini et al. 2009). The contribution of the oxide charges and the trap occupancy therefore has to be added on top of the interface state generation effect.

The Hole-Trapping Model

Although the RDD model corrects some of the issues with the RD model, it still cannot sufficiently explain the time-dependent recovery effect. Therefore, fast trapping and detrapping of holes in preexisting or NBTI-generated traps in the oxide has been introduced to explain the fast recovery effect. Interestingly, this possibility was already suggested by some of the very first papers on NBTI, including the first paper on the RD theory (Jeppson and Svensson 1977). However, hole trapping cannot account for the exceptionally long recovery times and the temperature dependence of stress and recovery. Therefore, a combination of interface state generation and hole trapping has been suggested (Parthasarathy 2006; Huard et al. 2007). Later, Grasser and Kaczer demonstrated how a unified field and temperature acceleration law can account for both the stress and recovery phase (Grasser and Kaczer 2009). This suggests either a single dominant mechanism or hole trapping and defect generation mechanisms that are actually tightly coupled.

In 2009, Ielmini et al. (2009) proposed a physical model, which describes the permanent and recoverable NBTI component under one unified framework including hole trapping/detrapping and thermally activated relaxation. The basic equation for the hole filling rate df/dt of one trap state at energy E_T is given by:

$$\frac{df}{dt} = \begin{cases} -f\frac{P_{\text{tun}}P_{\text{e}}P_{\text{SR,in}}}{\tau_0} + (1-f)\frac{P_{\text{tun}}P_{\text{SR,out}}}{\tau_0} & \text{if } E_{\text{t}} < E_{\text{F}} \\ -f\frac{P_{\text{tun}}P_{\text{SR,in}}}{\tau_0} + (1-f)\frac{P_{\text{tun}}P_{\text{e}}^{-1}P_{\text{SR,out}}}{\tau_0} & \text{if } E_{\text{t}} > E_{\text{F}} \end{cases} \tag{3.37}$$

The first and second term on the right-hand side of Eq. (3.37) represent the hole capture and emission rates, respectively. f is the trap state filling probability, E_F represents the Fermi level energy and τ_0 is the attempt time (around 1e–14 s). Further, P_{tun} is the tunneling probability and P_e is the probability for the hole to be excited. To overcome the limitation of hole trapping in accounting for the temperature activation of NBTI, local thermally activated structural relaxation (SR) is taken into account. In Eq. (3.37), $P_{SR,in}$ and $P_{SR,out}$ represent the thermal excitation of atoms accompanying the hole capture and hole emission, respectively. SR can also explain the breaking/fixing of Si–H bonds and therefore no other interface-state generation model such as the RD or RDD model is needed. The model presented by Ielmini et al. (2009) is therefore a better alternative to more conventional NBTI models such as the RD model, offering a more complete explanation for the typical NBTI-related effects. However, from Eq. (3.37) it is clear that this model again focuses on explaining the NBTI effect rather than modeling it for circuit simulation. The model not only describes the behavior of one oxide trap as opposed to the trapping behavior in an entire device. Further the model parameters are also not directly linked to common transistor design parameters, making model calibration and evaluation very hard.

Similar observations resulted from the work done by Grasser et al. (2011). They did measurements on very small devices ($W \times L < 100 \times 100$ nm) revealing recovery behavior in discrete steps. This is not consistent with a diffusion-limited process, but rather with the capture and emission of individual holes. The properties of these discrete steps were found to be fully consistent with charge trapping in the context of random telegraph noise (RTN) and 1/f noise. The difference between RTN and NBTI is explained as follows: RTN results from a limited number of defects trapping or detrapping charges within the experimental time window. Due to the strong bias dependence of the capture time constant, however, many more defects contribute to NBTI. NBTI is therefore considered as an non-equilibrium response of these defects, where RTN is their quasi-stationary behavior.

3.3.2 A BTI Compact Model for Circuit Simulation

The previous section has discussed the most well known NBTI models. Unfortunately, none of the presented models is suitable as a compact model for analog circuit simulation. Indeed, most models focus on *explaining* the NBTI phenomenon, rather than developing an all-inclusive compact model that is formulated in terms of transistor design parameters. Further, existing compact models only model the recovery effect for digital stress voltages (i.e. square waves) or only include the impact of fixed stress.

This section discusses the development, calibration and validation of an NBTI compact model, intended for analog and mixed-signal circuit reliability simulation (see Chap. 5). The proposed model is based on research suggesting that the NBTI phe-

nomenon consists of two tightly coupled mechanisms (Kaczer 2008; Grasser et al. 2009):

1. generation of defects close to the silicon/oxide interface resulting in a permanent component P, and
2. hole trapping in the gate oxide resulting in a recoverable component R.

The resulting model can handle both DC and AC stress and includes voltage and temperature scaling factors. The model is easy to calibrate featuring only 10 parameters, which can be extracted from a limited set of device measurements. Further, the proposed model can also be used to analyze the impact of PBTI, an effect similar to NBTI and typically observed in high-k metal-gate nFETs (Degraeve 2008). Finally, the model can be extended to predict stochastic BTI effects typically observed in ultra-scaled CMOS technologies (Kaczer et al. 2010). For analog circuits, however, the latter is less of an issue since matched transistors in analog circuits are typically very large which reduces the effect of time-dependent mismatch.

The NBTI phenomenon is linked to hole trapping in pre-existing oxide traps and the creation of defects at the silicon/oxide interface (Grasser et al. 2009). The former can be released when the gate voltage stress is lowered resulting in a recoverable component R, while the latter permanently captures holes resulting in a permanent degradation component P:

$$D = P(V_{\text{str}}, t_{\text{str}}) + R(V_{\text{str}}, t_{\text{str}}, t_{\text{r}}) \tag{3.38}$$

with D the total degradation, V_{str} the applied stress voltage (i.e. $V_{\text{str}} = |V_{\text{GS}} - V_{\text{TH}}|$), t_{str} the stress time and t_{r} the relaxation time (when a reduced stress voltage is applied).

The Recoverable NBTI Component

The recoverable component R is related to hole trapping and detrapping in pre-existing oxide traps and has a relaxation behavior for which the first derivative decays with $1/t_{\text{r}}$ over several decades (Grasser et al. 2009; Kaczer 2008):

$$\frac{d \log_{10}(\Delta V_{\text{TH,R}})}{dt_{\text{r}}} \propto 1/t_{\text{r}} \tag{3.39}$$

This behavior is similar to random telegraph noise (RTN), which is typically modelled with geometrically distributed RC-elements such as depicted in Fig. 3.7 (Kaczer 2008). For each RC-element the resistive value K is taken as constant, while the capacitor values are varied with a factor 10 from element to element.[1] At any moment in time, the recoverable NBTI component can be found as the sum of all the voltages $V_{\text{C}i}$ on each capacitor C_i. When the stress voltage V_{str} is varied in time, the voltage

[1] A base different from 10 can also be used and would alter the model parameters, but it would not affect the overall behavior of the model.

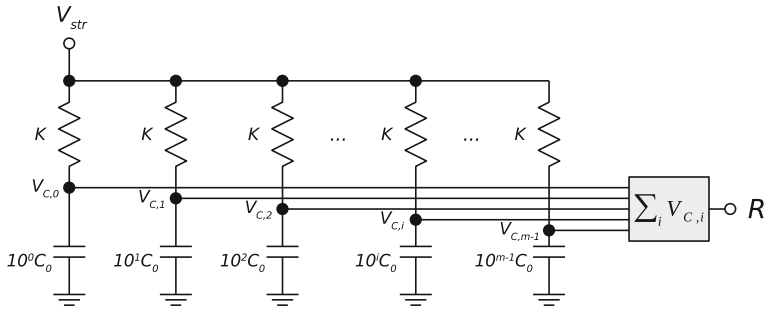


Fig. 3.7 A circuit with geometrically distributed RC-elements has a behavior similar to the stress and relaxation behavior of the recoverable NBTI component R (Kaczer 2008)

on each capacitor will also change (i.e. related to charge trapping or detrapping in the gate oxide) and this will result in an increase or decrease of R. However, the question remains how different values for the resistor K and the capacitors C_i relate to the actual stress and relaxation behavior of a transistor under NBTI stress. Fortunately, solving the differential equations of the RC-circuit and rewriting the result for a few limit cases results in a more comprehensible model for the recoverable NBTI component:

$$\begin{aligned} R &= \sum_{i=0}^{\infty} V_{\mathrm{C}i} \\ &= \sum_{i=0}^{\infty} \left(V_{\mathrm{str}} + (V_{\mathrm{C}i,0} - V_{\mathrm{str}}) \exp\left(\frac{-dt}{10^i K C_0} \right) \right) \end{aligned} \tag{3.40}$$

where $V_{\mathrm{C}i,0}$ is the initial voltage on every capacitor C_i (e.g. $V_{\mathrm{C}i,0} = 0$ for a fresh transistor), K is a constant and V_{str} is a fixed stress voltage, applied during time interval dt. When m represents the number of RC-elements in the system, Eq. (3.40) can be rewritten as:

$$R = m V_{\mathrm{str}} + \sum_{i=0}^{m-1} (V_{\mathrm{C}i,0} - V_{\mathrm{str}}) \exp\left(\frac{-dt}{10^i K C_0} \right) \tag{3.41}$$

$$= m V_{\mathrm{str}} + \sum_{i=0}^{m-1} (V_{\mathrm{C}i,0} - V_{\mathrm{str}}) \exp(-10^{\xi - i}) \tag{3.42}$$

$$= m V_{\mathrm{str}} + \sum_{k=\xi-(m-1)}^{\xi} (V_{\mathrm{C}i,0} - V_{\mathrm{str}}) \exp(-10^k) \tag{3.43}$$

with $\xi(dt) = \log_{10}(\frac{dt}{KC_0})$. The smallest time constant in the system is $\tau_0 = KC_0$, representing the oxide traps with the fastest trapping and detrapping rate. The traps with the slowest trapping rate are determined by $\tau_m = 10^{m-1}KC_0$. Typically $m = 25$ is sufficient for the actual implementation of the model, since it covers all trapping rates between 1e–15 s and 100 years. For limit cases $V_{\mathrm{C}i,0}$ is equal over all i:

1. for a fresh transistor: $V_{\mathrm{C}i,0} = 0,\ i = \{0, \ldots, m-1\}$
2. for a transistor stressed with voltage V_1 during $dt \gg \tau_m$: $V_{\mathrm{C}i,0} = V_1,\ i = \{0, \ldots, m-1\}$.

Equation (3.43) can, in those cases, be written as:

$$R = mV_{\mathrm{str}} + (V_{\mathrm{C},0} - V_{\mathrm{str}}) \sum_{k=\xi-(m-1)}^{\xi} \exp\left(-10^k\right) \tag{3.44}$$

$$\text{with } V_{\mathrm{C},0} \in \{0, V_1\}$$

Next, for $\xi \in \mathbb{N}$, the sum of exponential terms in Eq. (3.44) can be written as[2]:

$$\begin{aligned}
&\sum_{k=\xi-(m-1)}^{\xi} \exp\left(-10^k\right) \\
&= \underbrace{\exp\left(-10^{\xi-(m-1)}\right) + \cdots + \exp\left(-10^{-2}\right)}_{=m-2-\xi} + \\
&\underbrace{\exp\left(-10^{-1}\right) + \exp\left(-10^{0}\right)}_{=1.273} + \\
&\underbrace{\exp\left(-10^{1}\right) + \cdots + \exp\left(-10^{\xi}\right)}_{=0}
\end{aligned} \tag{3.45}$$

Combining Eqs. (3.44) and (3.45) then results in:

$$\begin{aligned}
R &= mV_{\mathrm{C},0} + (V_{\mathrm{C},0} - V_{\mathrm{str}})(-\xi - A) \\
&= mV_{\mathrm{C},0} + (V_{\mathrm{str}} - V_{\mathrm{C},0})(\log_{10}\left(\frac{t}{\tau_0}\right) + A)
\end{aligned} \tag{3.46}$$

with $\tau_0 = KC_0$ and $A \approx 0.73$. Also note how dt was replaced by t, indicating that the stress V_{str} is assumed to be constant over the entire measurement time. From Eq. (3.46) it is now clear that the slope of the equivalent RC-circuit has a linear dependence on the stress voltage V_{str}. This is however not in accordance

[2] One can proof that, when adding a small correction factor, Eq. (3.45) is also valid for $\xi \in \mathbb{R}$, but for the sake of simplicity, this is not included here.

with measurement results (i.e. in reality NBTI has a power-law stress dependence (Grasser et al. 2009)). Furthermore, the slope and offset of the forward reaction (i.e. hole trapping) can be different from the slope and offset of the reverse reaction (i.e. hole detrapping) (Kaczer et al. 2008; Grasser et al. 2009). Therefore Eq. (3.47) is proposed:

$$R = \begin{cases} m V_{\mathrm{C},0} + (V_{\mathrm{str}}^{n_\mathrm{V}} - V_{\mathrm{C},0}) \log_{10}\left(\dfrac{t^{n_\mathrm{F}}}{0.19\tau_\mathrm{F}}\right) & \text{if } V_{\mathrm{str}}^{n_\mathrm{V}} \geq V_{\mathrm{C},0} \\ m V_{\mathrm{C},0} + (V_{\mathrm{str}}^{n_\mathrm{V}} - V_{\mathrm{C},0}) \log_{10}\left(\dfrac{t^{n_\mathrm{R}}}{0.19\tau_\mathrm{R}}\right) & \text{if } V_{\mathrm{str}}^{n_\mathrm{V}} < V_{\mathrm{C},0} \end{cases} \tag{3.47}$$

where n_V is the voltage scaling factor and n_F and n_R are the slope for the forward and reverse reaction. τ_F and τ_R are the time constants for the forward and reverse reaction.[3] Equation (3.47) now represents a comprehensive model for the recoverable NBTI component R which is in full accordance with reality and only features 5 well-defined parameters: n_V, n_F, n_R, τ_F and τ_R. Equation (3.47) is only valid for limit cases (see Eq. (3.44)), but can be used to calibrate the model parameters for a specific CMOS process. Model evaluation, for a transistor subjected to an arbitrary time-dependent stress voltage, can be done with Algorithm 1.

Algorithm 1 Recoverable NBTI Component

INPUT: V_str, $V_{\mathrm{C}i}(t)$, t, Δt, $t_{\mathrm{F}i}$, $t_{\mathrm{R}i}$

for i=0:m-1 **do**
  Calculate $V_{\mathrm{C}i}(t+\Delta t)$:
  if $V_{\mathrm{C}i}(t) < V_{\mathrm{str}}^{n_\mathrm{V}}$ **then**
    Charge C_i (i.e. hole trapping):
    if $t_{\mathrm{F}i} < t_{\mathrm{R}i}$ **then**
      $t_{\mathrm{F}i} = t$
    end if
    $dt = (t + \Delta t - t_{\mathrm{F}i})^{n_\mathrm{F}} - (t - t_{\mathrm{F}i})^{n_\mathrm{F}}$
    $V_{\mathrm{C}i}(t+\Delta t) = V_{\mathrm{str}}^{n_\mathrm{V}} + (V_{\mathrm{C}i}(t) - V_{\mathrm{str}}^{n_\mathrm{V}})\exp\left(\frac{-dt}{10^i \tau_\mathrm{F}}\right)$
  else
    Discharge C_i (i.e. hole detrapping):
    if $t_{\mathrm{R}i} < t_{\mathrm{F}i}$ **then**
      $t_{\mathrm{R}i} = t$
    end if
    $dt = (t + \Delta t - t_{\mathrm{R}i})^{n_\mathrm{R}} - (t - t_{\mathrm{R}i})^{n_\mathrm{R}}$
    $V_{\mathrm{C}i}(t+\Delta t) = V_{\mathrm{str}}^{n_\mathrm{V}} + (V_{\mathrm{C}i}(t) - V_{\mathrm{str}}^{n_\mathrm{V}})\exp\left(\frac{-dt}{10^i \tau_\mathrm{R}}\right)$
  end if
end for
$R(t+\Delta t) = \sum_{i=0}^{m-1} V_{\mathrm{C}i}(t+\Delta t)$

OUTPUT: $R(t+\Delta t)$, $V_{\mathrm{C}i}(t+\Delta t)$, $t_{\mathrm{F}i}$, $t_{\mathrm{R}i}$

[3] The constant equal to 0.19 results from factor A in Eq. (3.46).

This algorithm is based on Eqs. (3.40) and (3.47) and takes as input: the stress voltage V_{str} (i.e. $V_{\text{str}} = |V_{\text{GS}} - V_{\text{TH}}|$) applied over stress time Δt, an array with the values of the voltages on each capacitor V_{Ci} at time t and two arrays with, for each capacitor, the last time point at which it was charged and discharged (i.e. $t_{\text{F}i}$ and $t_{\text{R}i}$ respectively).

The Permanent NBTI Component

The permanent NBTI component P is related to the creation of traps at the interface between the Si-channel and the gate oxide (Grasser et al. 2009). More recent publications argue how P could also result from the same underlying mechanism as the recoverable component R and that it is not really permanent but rather has a time constant outside the conventional measurement window (Grasser et al. 2011). However, even if that is the case, the proposed model still remains valid and the permanent component can be considered quasi-permanent within the lifetime of the circuit.

P has a rather small initial impact on the overall transistor degradation but, since P has a larger time exponent compared to R, its contribution to the overall degradation increases after longer stress times (Kaczer et al. 2008). The behavior of P can be modeled as:

$$P = C_{\text{P1}} \exp\left(C_{\text{P2}} V_{\text{str}}\right) t^{n_{\text{P}}} \tag{3.48}$$

with C_{P1}, C_{P2} and n_{P} technology-dependent parameters and V_{str} the applied stress voltage. Since the permanent part does not recover when the stress is reduced, the effect of time-varying stress can be calculated with (Parthasarathy 2006):

$$P_2 = \left[P_1^{1/n_{\text{P}}} + \left(C_{\text{P1}} \exp\left(C_{\text{P2}} V_{\text{str2}}\right)\right)^{1/n_{\text{P}}} (t_2 - t_1) \right]^{n_{\text{P}}} \tag{3.49}$$

where V_{str2} represents the stress voltage from t_1 till t_2. P_1 and P_2 represent the permanent NBTI component at t_1 and t_2 respectively. Ultimately, to model the permanent component P, only 3 parameters are required: C_{P1}, C_{P2} and n_{P}.

Temperature Scaling

From their experiments Grasser and Kaczer (2009) concluded how the temperature dependence of NBTI follows the Arrhenius law for both the permanent and the recoverable NBTI components:

$$D = (R + P) C_{\text{T}} \exp\left(-\frac{E_{\text{a}}}{kT}\right) \tag{3.50}$$

with the temperature T expressed in degrees Kelvin. Parameter k is the Boltzmann constant, E_a is the activation energy expressed in eV and C_T is a process-dependent constant. Parameters E_a and C_T need to be calibrated from measurements.

3.3.3 Model Calibration and Validation

Section 3.3.2 has proposed a compact model to simulate the impact of NBTI on analog circuits. In this section, the model is calibrated and verified in a 1.9 nm EOT SiON pFET technology. Given the importance of NBTI in nanometer CMOS processes, it is important to have an accurate compact model that can be characterized for any technology with only a limited set of simple measurements. This section first discusses the measurement setup. A clear calibration methodology is proposed. Finally, the extracted model parameters and the proposed model are verified with a time-varying stress voltage applied to a test transistor. The result is compared to a conventional DC-only model typically used for circuit reliability assessment.

Measurement Setup

Due to the extremely fast recovery behavior of NBTI, which occurs as soon as the stress condition of a transistor is changed (e.g. when interrupting the stress to measure the I-V curves), accurate evaluation of BTI degradation is very challenging. To cope with this problem, various measurement techniques have been proposed in literature:

1. **Measurement-stress-measurement (MSM)**: these conventional techniques consist of a sequence of stress periods in which a transistor is subjected to a stress voltage. After every stress period, the impact of the NBTI degradation on the behavior of the transistor is measured during a measurement phase. To reduce the impact of recovery, these techniques have been optimized by minimizing the delay between the removal of the stress and the first measurement point. Typically, either the drain current around the threshold voltage is recorded (Kaczer et al. 2005) or a threshold current is enforced through the device (Reisinger 2006). However, it is unclear whether the measurement delay can be made sufficiently short in order to measure true degradation characteristics.
2. **On-the-fly (OTF)**: these techniques monitor the degradation of the drain current in the linear regime, directly under stress conditions (Denais et al. 2004; Reisinger et al. 2007). This set of techniques thereby avoids any relaxation. On the other hand, it is difficult to record the initial state of the device under test (i.e. without stressing it involuntarily) and to extract the transistor parameters (e.g. threshold voltage).
3. **Charge-pumping (CP) and DCIV**: these allow to differentiate between oxide and interface charges (Neugroschel 2006; Mahapatra 2007). However, due to

the unclear interference between positive bias (for NBTI measurements) and relaxation effects, it is difficult to correctly interprets the measurement data.

A more elaborate comparison between the different techniques is given in (Grasser et al. 2008b). Overall, all measurement techniques used to determine the impact of NBTI on a transistor after it was stressed are prone to errors. For this work an MSM-based method has been adopted. In spite of the inherent delay, this method allows to accurately monitor the recovery behavior as a function of time, allowing the easy extraction of parameters for the compact model. Rapid, single-point measurements to evaluate the V_{TH}, right after stress, have been reported in literature (Kaczer et al. 2005). However, standard measurement equipment requires around 0.1 s to do a current measurement. At that point, the recovery process is already in progress. Extrapolation or model calibration based on such measurements is therefore meaningless. Further, even expensive, custom-built equipment allowing much faster measurements does not guarantee accurate results and is only available to a limited number of groups. For this work, a technique using off-the-shelf measurement equipment is used. The technique has been developed at IMEC and records a short recovery period during every measurement (Kaczer et al. 2008). From those measurements, these individual traces can be fitted together and the recoverable and permanent components of NBTI can be extracted independently.

The measurements have been performed using 1.9 nm-EOT SiON pFET devices with a $W/L = 1\,\mu m/1\,\mu m$ and a $V_{TH} = -0.137\,V@125\,°C$. The measurement equipment included two Keithley 2602 DMMs and a Süss PA300 probe station. A Perl script with nested LUA code was used to control the measurements. Figure 3.8 depicts a schematic representation of the setup. The device under test (DUT) was biased with a small drain-source voltage V_{DS} (i.e. around 50 mV); this to ensure that the voltage applied to the gate results in a uniformly distributed electric field over the gate oxide. V_{DS} remains constant during the entire measurement. The NBTI measurement sequence is depicted in Fig. 3.9 and proceeds as follows. First, an initial $I_{DS} - V_{GS}$-curve is recorded to obtain the performance of the transistor before stress. At this stage, it is important to apply sufficiently low voltages to avoid involuntary stressing

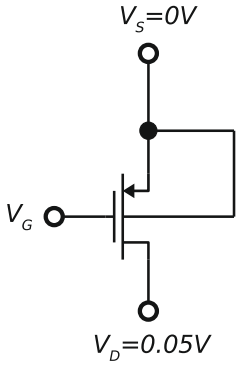


Fig. 3.8 Test setup for NBTI measurements

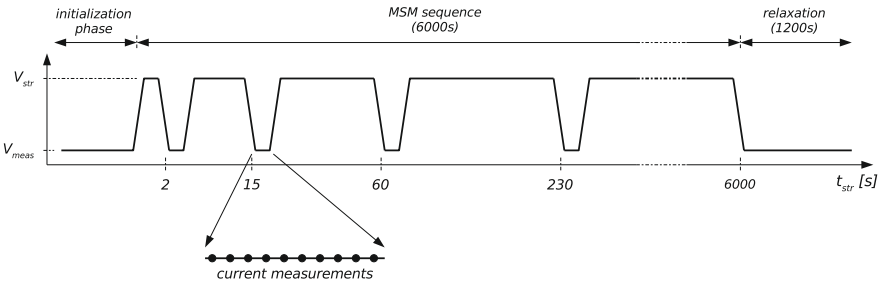


Fig. 3.9 NBTI measurement process. An initialization phase, where an $I_{DS} - V_{GS}$-curve is measured, is followed by a measurement-stress-measurement (MSM) sequence and finally a relaxation phase

of the device. Therefore, during this calibration phase, only voltages up to V_{meas} are applied. V_{meas} is chosen around the initial (before degradation) threshold voltage V_{TH0}. Then, the DUT is subjected to an MSM sequence with a total duration of 6000 s. During that sequence, stress phases ($V_{GS} = V_{str} + V_{TH0}$) are alternated with measurement phases ($V_{GS} = V_{meas}$). The duration of the stress phases is geometrically increased over time to cover several decades. The measurement phase following each stress phase has a duration of around 10 s, allowing to track the NBTI recovery behavior over a number of decades. In this work, $t_{str} = \{2, 15, 60, 230, 740, 2000, 6000\}$ s (also see Fig. 3.9). Eventually, after the MSM sequence, the transistor is relaxed during 1200 s. Figure 3.10 depicts an example set of measurement results resulting from the proposed method and for two different stress voltages. The transistor current is measured during the entire time, resulting in additional data during the stress phases. These data are equivalent to OTF measurements and can provide extra information about the behavior of the transistor. The geometrically distributed measurement points and associated recovery behavior are clearly visible. Further, the total accumulated stress time is significantly larger than the measurement time (i.e. the recovery time). Therefore, the impact of the recovery effect is fairly small and when reapplying stress after a measurement phase, the degradation almost instantly comes back to its original level.

Once the measurements are finished, the recorded transistor currents need to be converted into corresponding threshold voltages. During the measurement phase, the transistor operates in the subthreshold region ($V_{meas} \approx V_{TH0}$). Therefore the current is exponentially dependent on the applied voltage (Razavi 2001):

$$I_{DS} \propto \exp\left[\frac{q}{kT}\frac{(V_{GS} - V_{TH})}{n}\right] \tag{3.51}$$

with n a technology-dependent parameter ($n > 1$). Taking the logarithm of the current allows easy extraction of the threshold voltage change ΔV_{TH}:

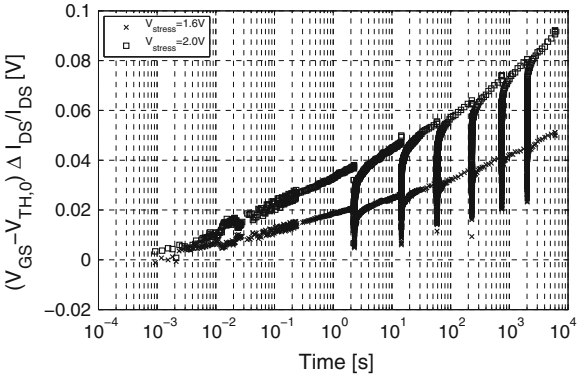


Fig. 3.10 NBTI-induced degradation measured for two stress voltages. At geometrically distributed time points, the stress is interrupted to extract information about the permanent and recoverable component. During stress, the current is also monitored, resulting in data similar to OTF measurements

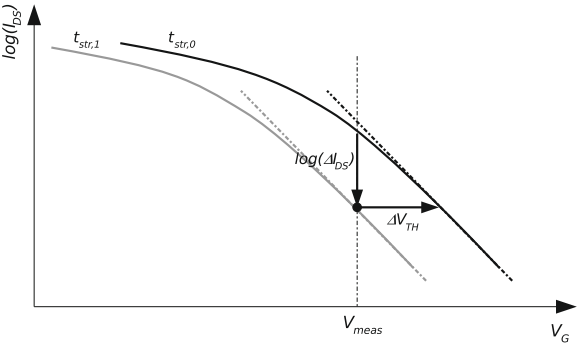


Fig. 3.11 To extract the NBTI-induced threshold voltage shift, the current is measured when the transistor is in the subthreshold region (i.e. *black dot*). Based on the initial $I_{DS} - V_{GS}$-curve (i.e. *black curve*), measured during the initialization phase, the shift in current and thus the corresponding shift in threshold voltage can be determined

$$\log(I_{DS} + \Delta I_{DS}) \propto \underbrace{\frac{q}{nkT}(V_{GS} - V_{TH})}_{\text{constant}} - \underbrace{\frac{q}{nkT}\Delta V_{TH}}_{\text{f(t)}} \tag{3.52}$$

This is also represented schematically in Fig. 3.11. For each measurement, the horizontal distance between the corresponding V_{GST} and the initial V_{GST0} is the shift in V_{TH} at that time point. In addition to an easy extraction of the V_{TH} shift, this method is also much less sensitive to mobility changes compared to measurements in the linear region.

Finally, the recoverable R and the permanent P NBTI component can be retrieved. During each measurement phase, part of the recovery behavior is traced. This

relaxation effect has a well defined log-like behavior that has been verified on different technologies and across multiple decades (Grasser 2007):

$$r(\zeta) = \frac{R(t_{\text{str}}, t_{\text{r}})}{R(t_{\text{str}}, 0)} = \frac{1}{1 + B\zeta^{\beta}} \tag{3.53}$$

where $R(t_{\text{str}}, 0)$ represents the value of the recoverable NBTI component right after stress, with stress time t_{str}. $R(t_{\text{str}}, t_{\text{r}})$ is the recoverable component after the device has been stressed and then recovered during time t_{r}. $r(\zeta)$ therefore represents the remaining fraction of the recoverable component during the relaxation period. $\zeta = t_{\text{r}}/t_{\text{str}}$ and is the universal relaxation time. B and β are parameters depending on the technology, the stress voltage and the temperature. For one stress condition (i.e. defined by the stress voltage and temperature), data is collected during 7 measurement phases (also see Fig. 3.9). This data can be used to extract values for the parameters B and β. In this work, this was done with software developed by the IMEC reliability research group. Eventually, the R and P component can be extracted for each stress time t_{str} point:

$$\Delta V_{\text{TH}} \approx P(t_{\text{str}}) + R(t_{\text{str}}, 0) \cdot r(\zeta) \tag{3.54}$$

Note how Eq. (3.53) is only valid for a fresh transistor that has been stressed with a fixed voltage and then relaxed. This equation therefore facilitates model parameter extraction and calibration but is not suited as a compact model for circuit reliability simulations.

Calibration

In the previous section, the NBTI measurement setup and procedure has been discussed. To calibrate the model the NBTI measurement was, in this work, repeated on different devices with various gate-source stress voltages:

$$V_{\text{GS}} \in \{-2.4\,\text{V}, -2.2\,\text{V}, -2.0\,\text{V}, -1.8\,\text{V}, -1.6\,\text{V}, -1.4\,\text{V}\} \tag{3.55}$$

Also, for every stress voltage, the measurement was repeated on 3 to 4 devices to estimate measurement errors and variations in the production process.

Figures 3.12 and 3.13 depict the recoverable component R, for various stress voltages and during the stress and relax phase respectively. Using those measurements and Eq. (3.47), the model parameters for the recoverable component are extracted: n_{V}, n_{F}, n_{R}, τ_{F} and τ_{R}. The measurement results for the permanent component P are plotted in Fig. 3.14. The model parameters for the permanent component, C_{P1}, C_{P2} and n_{P}, are extracted using Eq. (3.48). Figure 3.15 depicts the NBTI temperature dependence. The threshold voltage shift was normalized and plotted as a function of the inverse of the temperature. As predicted by Eq. (3.50), the NBTI effect follows an Arrhenius temperature dependence with activation energy E_{a} extracted from the

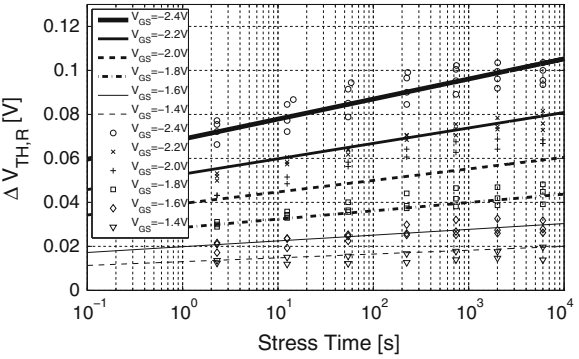


Fig. 3.12 The recoverable NBTI component (R) plotted as a function of the stress time for different stress voltages. The proposed NBTI model (*lines*) fits the measurement results (*markers*) very well

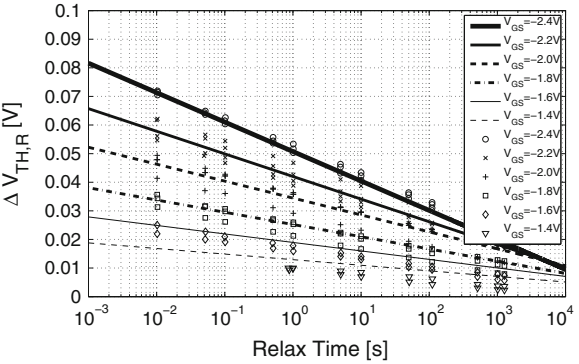


Fig. 3.13 The recoverable NBTI component (R) plotted as a function of the relax time for different stress voltages. The proposed NBTI model (*lines*) fits the measurement results (*markers*) very well

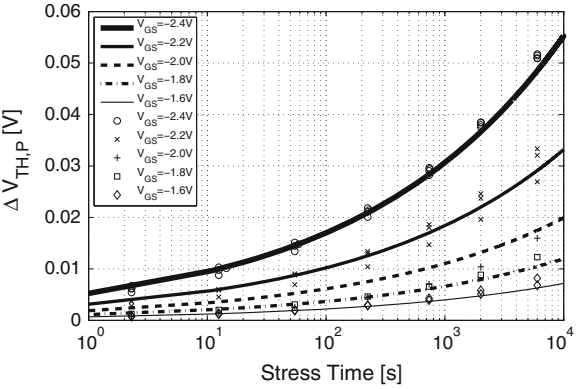


Fig. 3.14 The permanent NBTI component (P) plotted as a function of the stress time for different stress voltages. The proposed NBTI model (*lines*) fits the measurement results (*markers*) very well

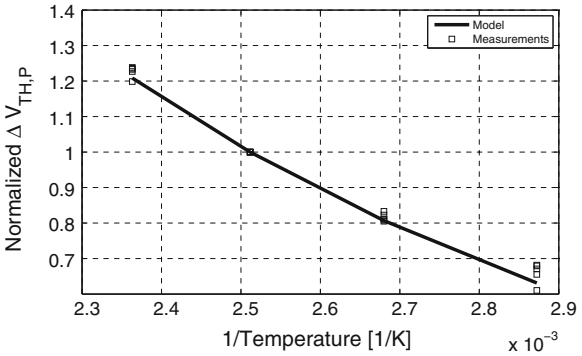


Fig. 3.15 The normalized threshold voltage shift as a function of the inverse of the temperature. The NBTI effect has an Arrhenius temperature dependence and the model fits the data very well

Table 3.2 Model parameters for a 1.9 nm EOT SiON CMOS technology (with the total degradation $D = \Delta V_{\mathrm{TH}}$)

R	$n_{\mathrm{V}} = 2.85$	
(13 samples)	$n_{\mathrm{F}} = 0.89$	$\sigma(n_{\mathrm{F}}) = 5.70\mathrm{e}{-2}$
	$n_{\mathrm{R}} = 1.00$	$\sigma(n_{\mathrm{R}}) = 3.00\mathrm{e}{-3}$
	$\tau_{\mathrm{F}} = 1.06\mathrm{e}{-6}$	$\sigma(\tau_{\mathrm{F}}) = 3.70\mathrm{e}{-7}$
	$\tau_{\mathrm{R}} = 2.06\mathrm{e}{-4}$	$\sigma(\tau_{\mathrm{R}}) = 2.00\mathrm{e}{-6}$
	$n_{\mathrm{P}} = 0.26$	
P	$C_{\mathrm{P1}} = 1.60\mathrm{e}{-2}$	$\sigma(C_{\mathrm{P1}}) = 2.50\mathrm{e}{-3}$
(13 samples)	$C_{\mathrm{P2}} = 2.55$	$\sigma(C_{\mathrm{P2}}) = 6.90\mathrm{e}{-2}$
Temperature	$C_{\mathrm{T}} = 2.06\mathrm{e}{-2}$	$\sigma(C_{\mathrm{T}}) = 5.55\mathrm{e}{-3}$
(4 samples)	$E_{\mathrm{a}} = 0.10$	$\sigma(E_{\mathrm{a}}) = 8.49\mathrm{e}{-3}$

measurements. Finally, Table 3.2 lists the extracted model parameters. The standard deviation on each model parameter is also given and results from parametric process variability and measurement errors.

Validation

To validate the proposed model, a triangle-shaped stress voltage was applied to the gate of a transistor (see Fig. 3.16). The measurement results and the corresponding evaluation of the proposed model are shown in Fig. 3.17. The model fits the measurement data very well. Figure 3.17 also shows the permanent and recoverable components P and R, as predicted by the model. To illustrate the importance of an NBTI model that can handle time-varying stress, such as the one presented here, Fig. 3.17 also depicts the threshold voltage shift when a fixed stress voltage (i.e. $V_{\mathrm{GS}} = -1.275$ and $V_{\mathrm{GS}} = -2.4$ V) would be evaluated instead of the actual stress input. Both clearly deviate from the actual degradation as measured. Using a sim-

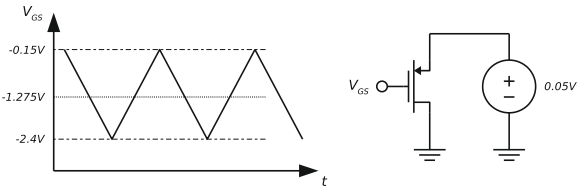


Fig. 3.16 The proposed model has been validated with a *triangle-shaped* stress voltage

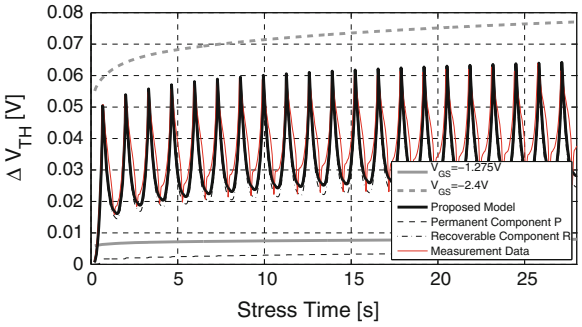


Fig. 3.17 Validation of the proposed model. The model matches the measurement data very well. Calculating the impact of applying only the average ($V_{GS} = -1.275\,V$) or the maximum ($V_{GS} = -2.4\,V$) input voltage is clearly too inaccurate

plified NBTI model for fixed stress, such as Eq. (3.29), combined with the average or maximum stress applied to the device is therefore too inaccurate to use in a circuit simulator. The model presented here solves that problem and allows accurate simulation of NBTI degradation under time-varying voltage stress.

3.3.4 BTI in Sub-45 nm CMOS

Section 3.3.2 has discussed the development of a compact model to evaluate the impact of NBTI on analog circuits. The model has been characterized and validated on a 65 nm CMOS technology. Scaling to even smaller technologies, however, reveals some additional problems:

1. Positive Bias Temperature Instability (PBTI) is observed in nMOS devices. This effect being negligible in SiON or SiO_2, large shifts have been reported in transistors processed in a high-k technology (Grasser 2010). The origin of the effect is believed to be similar to NBTI (Toledano 2011). Existing NBTI compact models such as presented above are therefore also suited for PBTI simulation, after they have been calibrated to measurements on nMOS devices.
2. In very small transistors, BTI has been observed to be a stochastic phenomenon with a ΔV_{TH} distribution due to individual charging and discharging events

(Kaczer 2010; Reisinger 2011) (also see Fig. 2.11 in Sect. 2.4). For these devices, BTI damage therefore has to be described as a stochastic process, rather than a deterministic shift. BTI can therefore result in time-dependent transistor mismatch. In Kaczer et al. (2010), discussed how the threshold voltage shift due to the trapping of a single charge follows an exponential distribution. These single-charge shifts can even exceed 30 mV for devices with $WL < 90 \times 35$ nm. The number of trapped charges is believed to be Poisson distributed. This allows to set up an analytical description of the total stochastic BTI threshold voltage shift for nanometer-size devices. For circuits with minimum-sized devices (e.g. in digital standard cells in sub-32 nm technologies) the number of traps per device can be as low as 10 or less. Reliability simulation therefore requires the individual monitoring of trapping and detrapping of all charges (Rodopoulos et al. 2011). Analog circuits, however, typically use larger transistors to reduce transistor mismatch due to process variations. For these larger transistors, it is sufficient to describe ΔV_{TH} with a distribution with a mean that corresponds to the value predicted by a deterministic model (see Sect. 3.3.2) and with a standard deviation equal to:

$$\sigma(\Delta V_{\mathrm{TH}}) = \frac{A_{\mathrm{BTI}}}{\sqrt{WL}} \Delta V_{\mathrm{TH}} \tag{3.56}$$

where A_{BTI} is a technology-dependent factor. Note how Eq. (3.56) has a similar size dependency compared to Pelgrom et al. (1989) law for transistor mismatch due to process variations. Equation (3.56) agrees with measurements, published in Agostinelli et al. (2004).

3.4 Time-Dependent Dielectric Breakdown

Dielectric breakdown (BD) results from oxide damage due to strong electric fields in nanometer CMOS technologies and gives rise to an increase of the transistor gate current. Different breakdown modes can be distinguished. Hard breakdown (HBD) results in a significant increase of the gate current and a loss of the gate voltage controllability of the device. Typically it is assumed that a hard breakdown in any transistor in the circuit results in circuit failure. For oxide thicknesses below 5 nm, Hard BD (HBD) can be preceded by SBD. SBD can be observed as a partial loss of the dielectric properties, resulting in a smaller increase of the gate current when compared to HBD. Multiple soft breakdowns do not necessarily result in a circuit failure.

For a circuit working at voltages within or even slightly above the nominal operating voltages, hard breakdown is very unlikely, while soft breakdown only results in a small increase of the gate current. Breakdown is therefore not the main focus of this work. Nevertheless, the performance of circuits such as power amplifiers, monolithic DC-DC converters and switched-capacitor circuits largely depends on the maximum voltage that can be applied on the gate and drain of a transistor. Here, accurate breakdown models can help to maximize the design margins while still guaranteeing

sufficient circuit lifetime. Therefore, for the sake of completeness, a brief overview of the most important breakdown models is given below. These models are however not used further in the text. First, hard breakdown modeling is discussed in Sect. 3.4.1, then modeling of soft breakdown is handled in Sect. 3.4.2.

3.4.1 Hard Breakdown

Hard breakdown dynamics have extensively been investigated in literature. As a consequence different models have been proposed in literature. The most important models are the thermochemical model, the anode-hole-injection model and the voltage-driven model.

The thermochemical model is also known as the E model, since it assumes a direct correlation between the applied electric field E_{ox} and the logarithm of the device lifetime. The model is explained by assuming the presence of weak chemical bounds in the oxide, which can break due to the applied electric field. The time-to-breakdown t_{BD} can be expressed as (McPherson and Baglee 1985):

$$t_{BD} \propto \exp(-\gamma E_{ox}) \exp\left(\frac{E_a}{kT}\right) \tag{3.57}$$

with γ the field acceleration factor ($\gamma \approx 1.1\text{decade/MV/cm}$) and E_a the thermal activation energy ($E_a \approx 0.6 - 0.9\,\text{eV}$). The E model shows a good fit when a long-term low-field voltage stress is applied, but is less accurate when predicting the impact of high-field stress.

The Anode-Hole-Injection Model is also referred to as the 1/E model and predicts a reciprocal electric-field dependence (Chen et al. 1985; Schuegraf and Hu 1994). Here, the breakdown phenomenon is explained as the trapping of holes at localized regions in the oxide:

$$t_{BD} \propto \exp(\beta / E_{ox}) \exp\left(\frac{E_a}{kT}\right) \tag{3.58}$$

with β a process-dependent electric field acceleration factor ($\beta \approx 350\,\text{MV/cm}$). In contradiction to the E model, the 1/E model has been proved to provide a good fit for data where a high electric field is applied. This model, however, is less accurate in low-field stress situations.

Each of the models reviewed above can only fit a limited range of the electric field. To alleviate this problem, researchers have proposed combined models (Hu and Lu 1999). Nevertheless, the applicability of these models does not appear to be valid for technologies with a gate-oxide thickness smaller than 5 nm (<250 nm). Breakdown of these ultra-thin oxides is shown to follow a power-law voltage dependence, rather than an exponential dependence (Wu et al. 2001). Also, the temperature dependence does no longer follow an Arrhenius temperature dependence (as for the E and 1/E

model) with an activation energy that increases with temperature. The voltage-driven model is expressed as:

$$t_{\mathrm{BD}} \propto V_{\mathrm{GS}}^{n(T)} \tag{3.59}$$

where the voltage acceleration factor n is temperature dependent. Due to this complex interaction between voltage, temperature and oxide time-to-breakdown, the modeling becomes much more complex than before. Later, this model was expanded to also include the area dependency and to model the distribution of the time-to-breakdown (Wu 2002; Wu and Su 2005). The latter is indeed required to do accurate TDDB circuit simulations, since Eqs. (3.57), (3.58) and (3.59) only predict the characteristic time-to-breakdown, while 63 % of the devices fail earlier than that. The time-to-breakdown follows a Weibull distribution and has a cumulative failure probability equal to:

$$F_{\mathrm{BD}} = 1 - \exp\left(-\frac{t}{t_{\mathrm{BD},63}}\right)^{\beta} \tag{3.60}$$

with $t_{\mathrm{BD},63}$ the characteristic lifetime (given by Eqs. (3.57), (3.58) and (3.59)) and β is the Weibull slope parameter. Combining Eqs. (3.59), (3.60) and the area and temperature dependency, one finally obtains a complete model for the time-to-breakdown in ultrathin oxides ($t_{\mathrm{ox}} < 5$ nm or <250 nm CMOS):

$$t_{\mathrm{BD}} \propto \left(\frac{1}{WL}\right)^{1/\beta} F_{\mathrm{BD}}^{1/\beta} V_{\mathrm{GS}}^{a+bT} \exp\left(\frac{c}{T} + \frac{d}{T^2}\right) \tag{3.61}$$

with W and L the transistor width and length respectively. The time-to-breakdown follows a non-Arrhenius model. Typical values for the model parameters (based on measurements on CMOS transistors with a t_{ox} ranging from 1.65 to 5 nm) are: $\beta = 1.64$, $F_{\mathrm{BD}} = 0.01\,\%$, $a = -78$, $b = 0.081$, $c = 8.81\mathrm{e}{-3}$ and $d = -7.75\mathrm{e}5$ (Li 2008). This model is to be used for breakdown predictions of circuits processed in advanced CMOS technologies. For older technologies (>0.25 µm CMOS) Eqs. (3.57) and (3.58) are more applicable.

3.4.2 Soft Breakdown

Since the introduction of thin-gate dielectrics ($t_{\mathrm{ox}} < 3$ nm), time-dependent dielectric breakdown received more attention in literature because of the presence of soft breakdown preceding hard breakdown. While hard breakdown results in a loss of the transistor gate-voltage controllability, soft breakdown does not necessarily coincide with device or circuit failure. After soft breakdown, a percolation cluster is created in the oxide. This results in a slight increase of the gate current. During the wearout (WO) phase, this percolation path wears out until hard breakdown occurs. This is indicated on Fig. 3.18 where t_{SBD} represents the soft breakdown time, t_{WO} is the

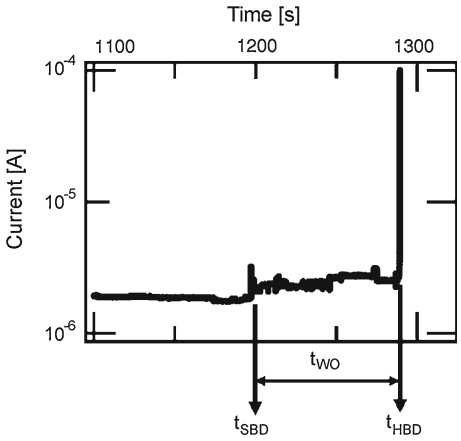


Fig. 3.18 Hard breakdown can be preceded by soft breakdown in gate oxides with a thickness smaller than 3 nm. The time between soft and hard breakdown is the wearout time (Sahhaf 2009)

wearout time and t_{HBD} is the hard breakdown time. Similar to hard breakdown, soft breakdown is also Weibull distributed and the time-to-breakdown has a power-law dependence. The wearout phase has a different behavior and is studied and modeled in (Sahhaf 2009). Most soft breakdown models published in literature only model the occurrence of one soft breakdown spot. In reality, however, multiple soft breakdowns can occur. Also, in contradiction to hard breakdown, the probability to have a circuit malfunction resulting from only one soft breakdown spot is very small. Therefore, it is important to model the number of (soft) breakdowns after a specific stress time, rather than the time to (the first) breakdown (as described by Eq. (3.61) for hard breakdown). The probability to have n SBD defects at time χ can be described with a Poisson distribution (Wu and Su 2005):

$$P_n(t) = \frac{\chi^n}{n!}\exp(-\chi) \tag{3.62}$$

$$\chi = \left(\frac{t}{t_{\mathrm{SBD},63}}\right)^{\beta}$$

$$t_{\mathrm{SBD},63} = t_{\mathrm{SBD,ref}}\left(\frac{WL}{A_{\mathrm{ref}}}\right)^{1/\beta}\left(\frac{V_{\mathrm{GS}}}{V_{\mathrm{ref}}}\right)^{\gamma}$$

where $t_{\mathrm{SBD,ref}}$ is the time-to-breakdown at the 63rd percentile for a reference transistor with area A_{ref} stressed at V_{ref}. β and γ are process-dependent parameters. Example values for the model parameters (extracted for a 0.9 nm EOT CMOS technology) are $\beta = 0.7$ and $\gamma = -62$ (Sahhaf 2009). Equation (3.62) is only valid for fixed stress voltages. However, while a circuit is aging, the transistor stress voltages might change due to aging-induced transistor parameters shifts. A dynamic SBD model, including support for changing operating points, is therefore required. To find such a model, the probability to have n SBD spots at time t_2, $P_n(t_2)$, can be looked at as the probability to achieve n' ($n \geq n'$) SBDs at time t_1, multiplied by the

probability to create an extra $n - n'$ breakdown spots between t_1 and t_2:

$$P_n(t_2) = \sum_{n'=0}^{\infty} \left[P_{n'}(t_1) \frac{\Delta\chi^{n-n'}}{(n-n')!} \exp(-\Delta\chi) \right] \tag{3.63}$$

$$\Delta\chi = \left(\frac{t_2 - t_1}{t_{\mathrm{SBD}} | V_{\mathrm{GS1}} = V_{\mathrm{GS2}}} \right)^{\beta}$$

with V_{GS1} the stress at t_1, V_{GS2} the stress at t_2 and $V_{\mathrm{GS,1}}$ and $V_{\mathrm{GS,2}}$ not necessarily the same. Equation (3.63) can now be used to calculate a PDF for the number of soft breakdowns after a specific stress waveform has been applied.

3.5 Aging-Equivalent Transistor Model

Sections 3.2.2, 3.3 and 3.4 have proposed a set of compact models for the most important transistor aging effects. These models translate circuit-related parameters such as stress voltages and transistor dimensions into an aging-induced change of the device behavior. At a physical level, this change is related to an increase in the number of interface and oxide traps. The latter directly affects the transistor threshold voltage, but changes also other transistor parameters such as the carrier mobility or the gate current. To simulate the impact of these changes on circuit functionality, an aging-equivalent transistor model is required that includes all these time-dependent changes.

Commercial simulation tools such as the reliability simulator integrated in Eldo (Mentor graphics 2012), allow to directly adjust device model parameters. This, however, requires a good understanding of the impact of transistor aging on the large number of parameters included in modern device models (e.g. BSIM4 contains around 220 parameters). Alternatively, one can add extra circuit elements such as voltage and current sources to include the impact of aging on the transistor. A good overview of the aging-inclusive transistor compact models is given in (Li et al. 2008). In this work, the aging-equivalent model depicted in Fig. 3.19 is used. The next sections discuss how to calculate the magnitude of each of the additional circuit elements shown in Fig. 3.19.

3.5.1 Threshold Voltage

The aging-induced threshold voltage shift is directly linked to the number of oxide and interface traps generated (Tam et al. 1984; Wang et al. 2007):

$$\Delta V_{\mathrm{TH}} = \frac{q(N_{\mathrm{IT}} + N_{\mathrm{OT}})}{C_{\mathrm{ox}}} \tag{3.64}$$

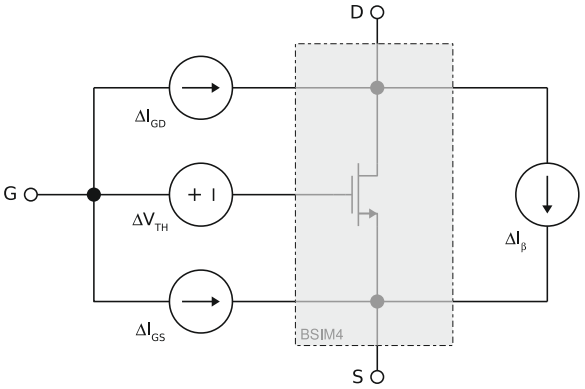


Fig. 3.19 The impact of aging on the performance of an nMOS transistor is modeled by adding extra circuit elements to the nominal device. A similar subcircuit can be used for a pMOS transistor

with C_{ox} the oxide capacitance per unit area. The change in threshold voltage shift is modeled with a voltage source ΔV_{TH} in Fig. 3.19.

3.5.2 Carrier Mobility

Sun and Plummer (1980) derived an empirical equation to describe the change in channel carrier mobility due to the formation of interface traps. The equation only applies to interface traps in the transistor inversion region. For HCI these traps are generated in the depleted pinch-off region of a transistor in saturation. Thus the effective mobility only decreases when the drain voltage is reduced and the pinch-off region decreases or even disappears. Similarly, in case of bias temperature instability, the carrier mobility shift is also larger for transistors operating in the linear region. Further, if an aging effect is dominated by oxide traps (e.g. BTI effects in sub-65 nm technologies (Grasser et al. 2009)), the carrier mobility shift will be rather small. Therefore:

$$\beta(t) = \frac{\beta_0}{(1 + a\gamma\Delta N_{\text{IT}})} \tag{3.65}$$

where α is a process-dependent parameter and β_0 is the current factor for a fresh transistor. γ is a parameter between 0 and 1 to indicate the fraction of the interface traps located in the inversion region. γ depends on V_{DS} but is close to zero when the transistor remains in saturation and becomes one as the transistor goes to the linear region. Parameter $a \approx 3$ for the IMEC 65 nm process used to also calibrate the HCI and NBTI models presented in the Sects. 3.2.2 and 3.3. To simulate the impact of the carrier mobility shift, an extra current source ΔI_β is added to the transistor model in Fig. 3.19. The value of this current source is:

$$\Delta I_\beta = \frac{\Delta\beta}{\beta_0} I_{\mathrm{DS0}} \tag{3.66}$$

with I_{DS0} the drain-source current of the transistor before it is aged and $\Delta\beta$ is the aging-induced change in current factor.

3.5.3 Oxide Breakdown

Oxide breakdown results in an increase of the gate current. As discussed in Sect. 3.4, this breakdown behavior is extremely complicated and the exact change in gate current depends on many parameters including the breakdown location, the transistor type, the applied gate voltage, the oxide area, etc. Further, in most analog circuits transistors are stressed with moderate voltages. Circuit failure due to breakdown is therefore very unlikely. To limit simulation time, the change in gate current resulting from the first soft breakdown over wearout to hard breakdown is therefore modeled with a first order model featuring two additional current sources (see ΔI_{GD} and ΔI_{GS} in Fig. 3.19).[4] These current sources follow a time-dependent behavior as proposed in (Li et al. 2008):

$$\Delta I_{\mathrm{GS}} = \alpha I_0 \left(\exp(\kappa \cdot e^{(\beta V_{\mathrm{GS}} - \theta t_{\mathrm{ox}})} \cdot t) - 1\right) \tag{3.67}$$

$$\Delta I_{\mathrm{GD}} = (1-\alpha) I_0 \left(\exp(\kappa \cdot e^{(\beta V_{\mathrm{GS}} - \theta t_{\mathrm{ox}})} \cdot t) - 1\right) \tag{3.68}$$

where κ is determined by the initial gate current I_0 and the gate current at which the circuit fails. Parameters β and θ depend on the technology, but example values for a 1.5 nm EOT CMOS technology are: $\beta = 8.64$ and $\theta = 7.6$ (Li et al. 2008). Parameter α varies between 0 and 1 and is determined by the breakdown location (i.e. $\alpha = 1$ if the breakdown spot is located near the source). The probability to have a breakdown at a given location is assumed to follow a uniform distribution.

3.6 Aging Model for Hand Calculations

The transistor aging compact models presented earlier in this chapter, are accurate and well suited for circuit reliability simulation. However, they are too complex for hand calculations. Nevertheless, when designing a circuit in an advanced CMOS technology, a simple first-order model can help a designer early in the design phase. For example, to compare the impact of aging on different circuit topologies or to

[4] When analyzing circuits with transistors that are stressed near or above the nominal supply voltage for an extended period of time, a more complex model has to be used. Such models have been discussed in Sect. 3.4.

Table 3.3 First-order transistor aging model useful for hand calculations (parameters given for a 32 nm CMOS process)

$\mu(V_{TH})$	$V_{TH0} + \Delta V_{TH}$
	$\lvert\Delta V_{TH}\rvert = V_{GST}^{\alpha}(C + n\log(t))$
$\sigma^2(V_{TH})$	$\frac{A_{VTH}^2}{2WL} + \frac{A_{BTI}^2 \lvert\Delta V_{TH}\rvert}{WL}$
pMOS	$\alpha = 2.45$, $C = 4.0\text{e–}2$, $n = 1.3\text{e–}2$, $A_{VTH} = 2.4\text{e–}9$, $A_{BTI} = 5.7\text{e–}9$
nMOS	$\alpha = 2.35$, $C = 4.5\text{e–}2$, $n = 1.5\text{e–}2$, $A_{VTH} = 2.4\text{e–}9$, $A_{BTI} = 5.7\text{e–}9$

find a good trade-off between circuit reliability and performance. Circuit reliability simulation then typically requires too much time. Therefore, in this section a first-order aging model for hand calculations is proposed.

Transistor aging is especially becoming a problem in sub-45 nm technologies. This is mainly due to the increasing oxide electric fields, the severely aggravated PBTI effect in nMOS transistors and the increased time-dependent variability due to stochastic aging effects in these technologies. The first-order model, proposed in Table 3.3, therefore focuses on these advanced technologies. The model allows designers to calculate the average threshold voltage and the variance on the threshold voltage as a function of time. Representative model parameter values for a 32 nm HKMG CMOS technology are extracted from literature (Degraeve et al. 2008; Pae et al. 2008; Lewyn et al. 2009; Cho et al. 2010; Kaczer et al. 2010; Pae et al. 2010). V_{GST}, in Table 3.3, represents the gate-source overdrive voltage. A transistor biased in the subthreshold region ($V_{GST} < 0$) does not age. The time t is expressed in seconds and the absolute value of the V_{TH} increases for both nMOS and pMOS transistors.

The model is derived from the NBTI model proposed in Sects. 3.3.2 and 3.3.4. PBTI and NBTI are believed to be the most dominant aging effects in nm CMOS technologies (Degraeve et al. 2008; Pae et al. 2010). The model assumes a constant stress applied to the transistor and a logarithmic time dependence. Calculating the impact of time-varying stress indeed requires a model as presented in Sect. 3.3.2 and is too complex for hand calculations.

3.7 Conclusions

This chapter has discussed the development of a set of compact models for evaluating the impact of the most important transistor aging effects. Each model has been optimized for the simulation of analog circuits:

- **Hot Carrier Injection (HCI)**: although less important in nanometer CMOS technologies, this effect can still be a problem for high-voltage applications such as power amplifiers and inductor-based oscillators). Therefore a new HCI compact model has been proposed. This model has been validated on a 65 nm technology.

- **Bias Temperature Instability (BTI)**: this is the most important aging effect for analog circuit reliability in nanometer CMOS. Partial recovery of the transistor damage when the stress voltage is reduced significantly complicates modeling this phenomenon. A new NBTI compact model, that includes this recovery effect has been proposed. The model has been characterized and validated on a 65nm CMOS technology.
- **Time-Dependent Dielectric Breakdown (TDDB)**: although this phenomenon can cause abrupt circuit failure, the voltages at which this happens are typically not used for analog circuits in nanometer CMOS. Nevertheless, a set of compact models has been given. These models capture the breakdown effect from the first soft breakdown spot over the wearout of the oxide to the eventual hard breakdown event.

Finally, the last section in this chapter has presented an aging-equivalent transistor model. Extra current and voltage sources, added to the standard transistor BSIM model, emulate the impact of aging on the behavior of the transistor. Also, a simple first-order model useful for reliability hand calculations has been proposed. Both the aging-equivalent transistor model as well as the model for hand calculations will be used for circuit reliability analysis in future chapters.

Chapter 4
Background on IC Reliability Simulation

4.1 Introduction

Since the introduction of the first SPICE simulator in 1973 (Nagel and Pederson 1973), circuit designers use simulators to predict and optimize circuit performance at design time. This results in huge savings in development costs and enables a designer to maximize the performance of his or her circuit in a particular technology. Over time, computer-aided design (CAD) software has become more complex and more and more aspects related to IC development have been modeled and included in circuit simulation tools. As designers try to push their designs to the limit, using technologies with ever-smaller feature sizes, more and more reliability problems pop up. Guaranteeing sufficient product yield under the presence of process variations, for example, have become one of the first major IC reliability issues. To estimate the impact of process variations at design time, EDA companies have started to offer variation-aware simulation methods such as corner simulations or Monte-Carlo (MC) simulations.

With transistor aging effects having an ever increasing impact on circuit performance (see Chap. 2), circuit reliability simulation is another important part of a modern IC design flow. Without such a tool, designers are forced to use design margins, extracted from measurements on individual transistors (e.g. limit V_{DD} such that $\Delta V_{TH} < 50\,\text{mV}$ after 10 years). However, these margins are often too restrictive and can result in huge circuit overdesign. Also, these rules do not guarantee circuit reliability, especially for analog circuits which tend to be very sensitive to small transistor parameter variations. Accurate reliability simulation therefore enables a designer to significantly increase the circuit design space, to meet tougher circuit specifications and to guarantee reliable circuit operation.

In Sect. 4.2, this chapter first discusses the most important simulation methods published in literature. Special attention is given to BERT, which is one of the first reliability simulation toolsets, developed in the early 1990s by a group in UC Berkeley (Tu et al. 1993). Section 4.3 reviews the reliability simulators integrated in each of the

E. Maricau and G. Gielen, *Analog IC Reliability in Nanometer CMOS*,
Analog Circuits and Signal Processing, DOI: 10.1007/978-1-4614-6163-0_4,

major commercial SPICE simulators: the Mentor Graphics Eldo reliability simulator, Cadence RelXpert and Synopsys MOSRA. Since most of these simulators are based on the methods discussed in Sect. 4.2, the focus of Sect. 4.3 is on the completeness of functionality and the usability of the tools.[1] The advantages and disadvantages of the various academic and commercial simulators are discussed in Sect. 4.4. Finally Sect. 4.5 gives the conclusions of this chapter.

4.2 Literature Overview

Guaranteeing circuit reliability in the presence of transistor aging has been a problem since the mid 1980s. As a consequence, over the years many circuit reliability simulators have been proposed in literature. First, during the late 1980s and early 1990s, when hot carrier degradation was a major problem, tools such as HOTRON (Aur et al. 1987), RELY (Sheu et al. 1989) and BERT(CAS) (Tu et al. 1993) were developed. Later, when effects such as TDDB and NBTI became more important, a second group of simulation methods was proposed (Xuan et al. 2003; Parthasarathy 2006; Bestory et al. 2007; Yan et al. 2009; Wang et al. 2010). Below, the most important reliability simulation methods are reviewed in more detail.

4.2.1 Berkeley Reliability Tools (BERT)

The Berkeley Reliability Tools (BERT) are a set of methods, developed by Hu et al. at the University of California Berkeley in the early 1990s (Tu et al. 1993). The toolset allows to simulate the impact of hot-electron degradation in MOSFETs and bipolar transistors. Further, prediction of circuit failure due to oxide breakdown or electromigration in CMOS, bipolar and BiCMOS is supported. The toolset is written around a commercial circuit simulator such as SPICE. Although rather old, a lot of modeling and simulation elements used in BERT are still applied in modern commercial reliability simulators. As a main disadvantage, BERT assumes each aging effect to be independent from one another. However, in reality this is often not a valid assumption. For example, oxide traps generated due to hot carriers, will affect the oxide wearout degradation rate and can result in an early breakdown-related circuit failure.

As an input, the user must provide a netlist with the description of the circuit and technology-specific device model parameters for each aging effect. BERT obtains voltage waveforms for each transistor using a commercial circuit simulator. BERT itself is designed as a pre- and post-processor combination around the SPICE

[1] In accordance with the focus of this work, this chapter focuses on tools intended for analog circuit lifetime analysis. Other useful and complementary tools such as the Berkeley Design Automation AFS platform, the Solido Analog$^+$ Suite, the Muneda WiCkeD toolset or any toolset for digital circuit simulation and verification are not discussed here.

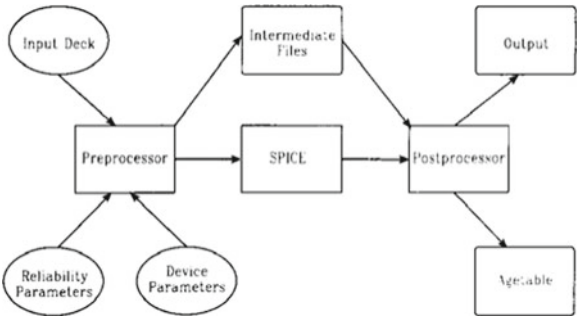


Fig. 4.1 The Berkeley reliability tools (BERT) consist of a pre- and post-processor around a SPICE engine (Tu et al. 1993)

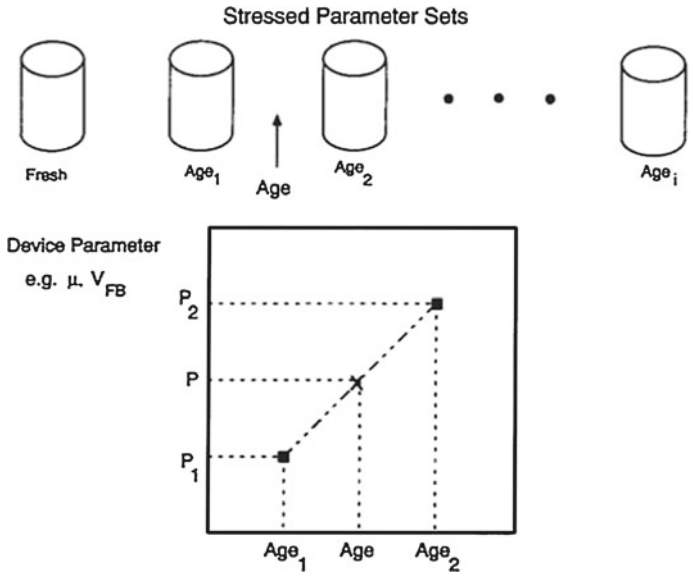


Fig. 4.2 Calculation of aged device parameters from pre-stressed devices and simulated AGE parameters. The *barrels* represent pre-stressed devices for specific AGE parameter values (AGE_1 to AGE_i). The degradation of a specific transistor in a circuit is represented by AGE. The graph illustrates how measured device parameter shifts are interpolated to find the device parameter shift corresponding to the degraded transistor in the circuit under test (Tu et al. 1993)

engine (see Fig. 4.1). Both processors each contain a module for each reliability phenomenon:

1. The circuit aging simulator (CAS) module simulates hot-carrier degradation (Fig. 4.2).
2. The circuit oxide reliability simulator (CORS) module simulates time-dependent dielectric breakdown.

3. The electromigration (EM) module simulates electromigration.
4. The bipolar circuit aging simulator (BiCAS) simulates hot-carrier degradation in bipolar transistors.

This modular structure allows easy expansion of the toolset, for example with a module to simulate the impact of radiation effects. The two most important modules with respect to the focus of this work are briefly explained in the following two paragraphs.

The BERT CAS (circuit aging simulator) module requires the circuit simulator to output the voltage waveforms at the terminals of all MOS transistors. The post-processor uses these waveforms to calculate the degradation rate of each transistor. This is done through an AGE parameter, which quantifies the amount of hot carrier degradation in each device as a function of the bias conditions and time. The AGE parameter for an nMOS transistor, for example, is based on the lucky-electron model (see Sect. 3.2.1 and Hu et al. 1985):

$$\mathrm{AGE} = \int_0^{T_{\mathrm{str}}} \frac{I_{\mathrm{DS}}}{W H_{\mathrm{HCI}}} \left(\frac{I_{\mathrm{sub}}}{I_{\mathrm{DS}}} \right)^{m_{\mathrm{HCI}}} dt \tag{4.1}$$

where W represents the device width and H_{HCI} and m_{HCI} are technology-dependent parameters. I_{sub} is the substrate current and I_{DS} is the drain current. T_{str} is the stress time. The amount of damage done to transistor model parameters such as the threshold voltage or the carrier mobility is then related to this AGE parameter:

$$\Delta V_{\mathrm{TH}} = f(\mathrm{AGE}) \tag{4.2}$$

The post-processor calculates the AGE parameter for each transistor and stores the result in an age table. The pre-processor then uses this age table to generate new transistor model parameters for each stressed device. This is done by interpolation between user-supplied process files. These files are based on stress measurements on actual transistors and contain shifts in device parameters as a function of the corresponding AGE parameter values. Finally, the user can simulate the modified netlist to analyze the impact of hot carrier degradation on the circuit under test. The BERT CAS module has been integrated in Cadence (see Sect. 4.3.2).

The BERT CORS (circuit oxide reliability simulator) module calculates the probability of circuit failure due to oxide breakdown. BERT uses the 1/E-model (see Sect. 3.4.1) to model the TDDB phenomenon. However, TDDB is a stochastic phenomenon and the time-to-breakdown is a statistical variable. Therefore, the CORS module calculates the probability for a specific device to break down before the user-defined circuit stress time. First, the 1/E-model is used to calculate the maximum effective oxide thickness $t_{\mathrm{ox,max}}$ that will cause a breakdown before the pre-defined circuit stress time. Then, a user-defined cumulative distribution function for the time-to-breakdown, corresponding to $t_{\mathrm{ox,max}}$, is used to calculate the probability that at least one device in the circuit fails.

Being one of the earliest and most complete simulators in its time, BERT had a big influence on the development of commercial reliability simulators. However, other reliability simulation methods have also been presented, not only during the same decade BERT was developed, but especially later when NBTI became one of the most important aging phenomena. The next section highlights some of these methods.

4.2.2 Other Reliability Simulators

Most circuit reliability simulators, such as the one discussed in the previous paragraph, are built around a SPICE simulator. As a consequence, such a simulator is independent of the SPICE engine which can be an advantage for designers working in a multi-simulator environment but also for the tool developers who can build their tool around existing simulators. The entire aging calculation, however, happens after the SPICE simulation and therefore requires extra time. Also, the SPICE simulator needs to store the waveforms on every node in the circuit, requiring memory and slowing down the SPICE simulation itself. In order to solve these problems, Parthasarathy proposed a method that is integrated in the SPICE simulator (Parthasarathy 2006; Parthasarathy et al. 2006). In this case, the reliability simulation is done while the transient simulation is running: i.e. the aging of each transistor is calculated at each transient step when the operating points are computed. Therefore, the simulator does not need to store all internal voltages and there is no extra time required to calculate the circuit degradation after the SPICE simulation. At the end of the SPICE simulation, the degradation is extrapolated to the desired circuit lifetime. Then, the circuit is evaluated again to extract the aged circuit performance. The main disadvantage of this method is that it is very simulator specific. This approach has been integrated in the Eldo reliability simulator (see Sect. 4.3.1).

Even if an integrated simulation approach such as described above is used, reliability simulation still requires a huge computational effort. To further reduce simulation times, alternative approaches have been explored. Bestory et al. (2007) proposed to use a hierarchical approach where each system sub-circuit is replaced by a behavioral model. Such a model not only includes input parameters and environmental parameters, but also supports transistor aging. Using a model to evaluate each sub-circuit in a system significantly speeds up reliability simulation and even allows to do a MC analysis on top of the reliability simulation. Then, one can calculate the circuit failure rate, as opposed to the performance of only one circuit sample. The main disadvantage of the work presented in Bestory et al. (2007) is the way how the sub-block models are constructed: i.e. using VHDL-AMS and designer knowledge. Indeed, in that case, the designer is required to know how to model the behavior of the circuit and also how to model the impact of transistor aging. This requires a huge effort and substantial knowledge about transistor aging effects.

4.3 Commercial Reliability Simulators

Above, the most important reliability simulation methods published in literature have been discussed. Some of these methods have been implemented in commercial products and are available for IC designers to evaluate the lifetime of their circuits. This section reviews each of the major commercial reliability simulators available on the market.

4.3.1 The Mentor Graphics Reliability Simulator

The reliability simulator provided by Mentor Graphics is integrated in Eldo, a SPICE circuit simulator (Mentor graphics 2012). The simulator is based on work done by Parthasarathy et al. (Parthasarathy 2006; Parthasarathy et al. 2006) and is intended to provide information about circuit performance shifts due to gradual transistor aging effects. Abrupt effects such as dielectric breakdown are therefore not supported. Further, this tool also does not work for the reliability simulation of circuits in sub-45 nm CMOS. Indeed, due to the limited number of dopant atoms in these technologies (<100), aging effects such as BTI that used to change gradually in older technologies, also change in discrete steps in these newer processes (see also Sects. 2.4 and 3.3). The simulator framework consists of two parts, which are briefly explained in the next sections:

1. A user-defined reliability model (UDRM) interface enabling users to implement their own reliability models.
2. A reliability simulator to analyse the impact of aging phenomena, modeled through the UDRM interface, on the behavior of a circuit.

The user-defined reliability model (UDRM) consists of a set of functions written in the C language. These functions can be used to define one's own device reliability compact model equations, such as the models presented in Chap. 3. These models are then evaluated by the simulation kernel during the reliability simulation. So-called interface functions can be used to extract circuit-specific information such as voltages, currents and sizes. This information is needed to accurately calculate the impact of the aging of each device in the circuit. Each transistor aging model should consist of two parts. The first part defines how to calculate a stress parameter. This is a time-independent quantity representing the stress on any device at a specific point in time. Then, a second set of equations defines how to relate this stress parameter to a time-dependent change in the device model parameters (i.e. BSIM parameters). This relationship can be different for different model parameters. Except for a first-order example model for hot carrier effects, the Mentor reliability simulator does not provide transistor aging models. Therefore, they must be defined by the user or provided by the foundry before a reliability simulation can be carried out.

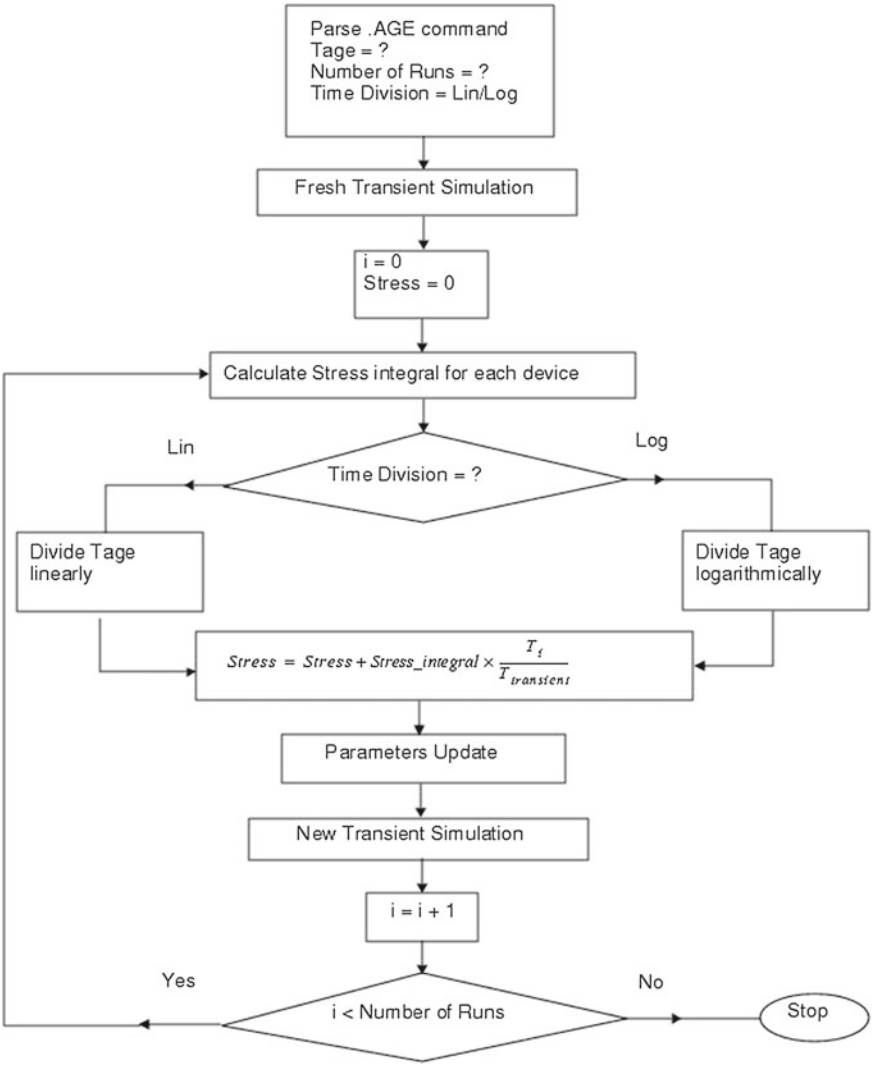


Fig. 4.3 Flow chart of the reliability simulation flow implemented in Eldo

The Eldo reliability simulator requires a normal netlist at the input. The netlist should contain a .TRAN (transient analysis) or .SST (steady state analysis) statement to allow extraction of the exact stress voltage on each node of each transistor in the circuit. This is important since most aging effects have a nonlinear dependence on the stress voltage (also see Chaps. 2 and 3). Calculating only the average DC stress voltage (e.g. with a DC operating point analysis) therefore does not yield a correct result. Figure 4.3 depicts a schematic representation of the reliability simulation flow implemented in Eldo. First, Eldo requires a .AGE command in the netlist to determine:

- The total circuit stress time (*Tage*),
- the number of time intervals *Tage* will be divided into (number of runs),
- the method with which to divide Tage: linearly or logarithmically.

Dividing the total stress time in a larger number of intervals will result in a more accurate aging simulation. Also, this will account for the gradual change in bias conditions as a result of the transistor aging. After parsing the .AGE command, a transient simulation on the fresh netlist is carried out. Then, the stress of every transistor is calculated, using the equations given in the user-defined reliability model (UDRM). All the calculated stress values are combined in a stress vector *Stress_integral*. Next, T_i is calculated to determine the next time point at which the aged simulation will be carried out. If linear time division is specified, then equal time divisions will be considered. However, if a logarithmic time division method is specified, the time divisions will be crowded at the beginning and get larger as *Tage* is approached. This can be useful since some aging effects such as HCI and BTI have a power-law time dependence. The calculated stress vector *Stress_integral* is then multiplied by a factor $T_i / T_{\text{transient}}$, where $T_{\text{transient}}$ represents the total user-defined transient simulation time. Here it is assumed that during T_i the aging-induced change in bias conditions is negligible. The extrapolated stress vector *Stress* is then added to the previous stress vector values (the impact of stress is calculated cumulatively through the runs). Next, the new stress vector is used to update the parameters of the model card for each ageable device in the circuit. Again this relationship is defined in the reliability model equations. Once this is done, the circuit is ready for a new aged transient simulation. This process is repeated until $t = Tage$, where a final transient simulation is done and the degraded performance of the circuit can be examined. The Eldo reliability simulator also offers some analysis options, such as a sensitivity analysis to identify which devices are sensitive and must not be stressed too much to avoid circuit failure.

4.3.2 The Cadence Reliability Simulator (BERT/RelXpert)

Cadence supports reliability simulation in Virtuoso Ultrasim and in the Analog Design Environment (ADE) (Cadence 2012). The reliability simulator offers simulation and analysis of the impact of gradual aging effects such as HCI and NBTI (also see Chap. 2). Therefore, the tool has similar capabilities and limitations compared to the reliability simulator included in Mentor Graphics Eldo (see Sect. 4.3.1).

Cadence supports two methods to model aging effects for a specific technology:

1. A table model (Aged Model): aged SPICE model parameters are extracted from a fresh device at a number of stress intervals. These model parameters form a set of aged model files. During the reliability simulation, the aging for each transistor in the circuit is calculated based on interpolation or regression of the values in these files. This approach is based on BERT (see Sect. 4.2.1 and Tu et al. (1993)).
2. An analytical model (AgeMOS): An analytical model describing each aging effect. This model should be provided by IC manufacturers, but users can also

develop their own model description. Such a model describes the change of a device model parameter as a function of the transistor age, process-specific AgeMOS parameters and the applied voltage or current. This approach is based on RelXpert, a tool developed by Celestry, and acquired by Cadence in 2003 (Celestry RelXpert 2012).

The analytical method (AgeMOS) allows a more accurate, more consistent and faster reliability simulation, but is also harder to develop. IC manufacturers do not always provide aging models and even if they do, the models often only include DC stress effects (e.g. the BTI recovery effect, as described in Sects. 2.4 and 3.3, is often not supported). A table model, on the other hand, is easier to construct, but is in most cases too inaccurate due to large extrapolation errors. Cadence therefore advises to use the AgeMOS model and even provides a basic HCI and NBTI analytical model (model parameters still need to be provided by the foundry).

To use the reliability simulation with Virtuoso Ultrasim, a transient analysis is required. This allows the simulator to extract accurate information about the exact stress on each device in the circuit. Further the user needs to add additional control statements to the Spectre SPICE netlist file. These include:

- A *.age* statement to specify the total stress time and intermediate points at which the degraded netlist is evaluated,
- A statement to specify the age calculation method (table model or analytical model),
- A *.deltad* statement, which enables a circuit lifetime calculation. The argument given with this statement tells the simulator the maximum relative parameter shift for any transistor in the circuit. For example, when *.deltad 10* is specified and the relative threshold voltage shift of an arbitrary transistor surpasses 10 % after a total stress time T_{str}, the circuit is considered to fail and the simulator returns a circuit lifetime equal to T_{str}.

The reliability simulator (using the more advanced AgeMOS analytical aging models) is also supported by the analog design environment (ADE). Under the hood, the reliability simulator in ADE, uses the same framework as the simulator in Virtuoso Ultrasim. Therefore similar simulator modes are offered, but specified through a visual interface rather than a text-based approach (also see Fig. 4.4).

4.3.3 The Synopsys Reliability Simulator (MOSRA)

Synopsis MOS reliability analysis (MOSRA) is a reliability simulator included in HSPICE and CustomSim (Synopsys 2012). MOSRA enables the analysis of the impact of HCI and BTI effects on integrated circuits. Again, only gradual aging effects are supported.

The Synopsys reliability simulation flow supports the use of custom models Pdeveloped by device modeling teams within a company or by foundries. Additionally,

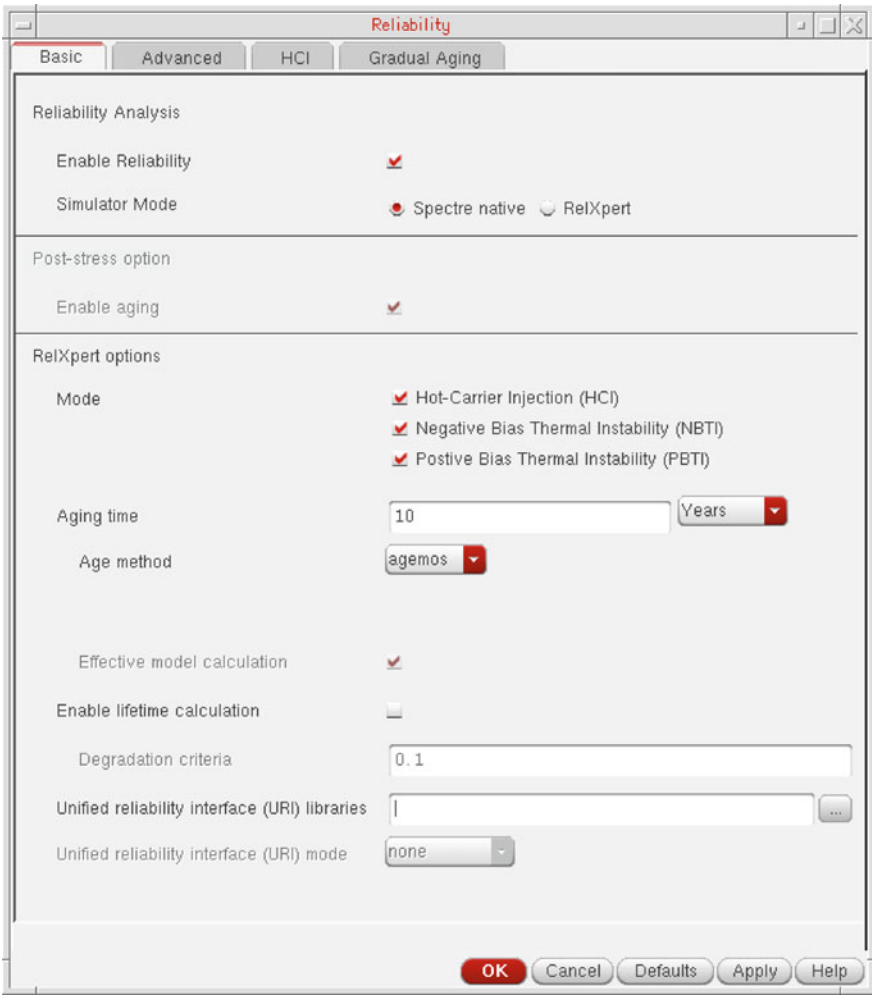


Fig. 4.4 The Cadence analog design environment (ADE) includes a reliability simulator to analyse the impact of gradual aging effects on the performance of a circuit. This simulator is also supported by Virtuoso Ultrasim, but requires additional statements to be included in the netlist file rather than through a visual interface

the tool comes with a set of compact models for both HCI and BTI degradation. Both models are fairly accurate, when compared to the aging models offered by Eldo or Cadence. The BTI model includes a term for the partial recovery effect that is essential for BTI. Unfortunately, only the duty cycle of the input is included in this calculation. Therefore, the model is actually only useful for the reliability simulation of digital circuits. The HCI model not only supports the well-known lucky-electron model (see Sect. 3.2.1) but also includes an extra term for accurate HCI simulation in the

Table 4.1 Circuit reliability simulators compared

	Name	HCI	BTI	TDDB	EM	PVT	AMS	VLSI
Aur et al. (1987)	HOTRON	x					x	
Sheu et al. (1989)	RELY	x						x
Tu et al. (1993)	BERT	x		x	x		x	
Xuan et al. (2003)	ARET	x			x			x
Parthasarathy (2006)	–	x	x				x	
Bestory et al. (2007)	–	x	x			x	x	
Yan et al. (2009)	–	x	x	x			x	
Wang et al. (2010)	SyRA	x	x					x
Mentor	Eldo RS	x	x				x	
Cadence	RelXpert	x	x				x	
Synopsys	MOSRA	x	x				x	

(Versions available on March 2012)

high-current regime. Model parameters need to be extracted from measurements or provided by the foundry.

The MOSRA simulation flow includes two phases: a pre-stress simulation phase and a post-stress simulation phase. During the pre-stress simulation phase, the simulator computes the electrical stress of user-selected transistors in the circuit. This calculation is based on the MOSRA aging models. The result is then extrapolated to calculate the total stress after a user-specified total circuit stress time. During the post-stress phase, a second simulation is launched. The degradation of the device characteristics is then translated to performance degradation at the circuit level. During the aging simulation, the tool also considers the impact of accumulated stress and therefore gradual changes of circuit bias conditions are also taken into account.

4.4 Discussion

Each of the techniques described above has certain advantages and disadvantages. Commercial reliability simulators are mostly based on methods or tools published in literature and therefore also inherit the same advantages and disadvantages. Table 4.1 lists the most important tools, published in literature and commercially available.

Older tools clearly focus on evaluating the impact of hot carrier degradation. Around 2005, when nitrogen was introduced in the gate stack, NBTI became increasingly more important. Therefore, each tool introduced after 2005 also includes the impact of (N)BTI in the reliability simulation. Only a very limited number of tools provide support for TDDB calculations. And even then, the support is limited. BERT (Tu et al. 1993), only calculates the time-to-first breakdown which does not necessarily coincide with an actual circuit failure. Yan et al. (2009) include the impact of multiple soft breakdowns and hard breakdown, but only the nominal behavior is

modeled. The distinctive statistical nature of TDDB is thus not taken into account (i.e. each transistor degrades in the same way and fails at the same time). Most tools also do not support reliability analysis including process variations (indicated PVT in Table 4.1). Nevertheless, as a designer it is important to consider this since one circuit could age faster than another, resulting in a dispersion of the lifetime (Bestory et al. 2007).

Most tools are intended for the simulation of analog or mixed-signal circuits. Typically, a transient simulation is required to extract the stress waveforms on every node of every transistor in the circuit. Then, after calculation of the degradation of each individual transistor, at least one other SPICE analysis is needed to evaluate the impact of aging on the circuit behavior. This already results in a fairly large computational effort and even more simulations are required in order to simulate the gradual aging-induced change of the circuit bias conditions. Although there are some differences in implementation (e.g. reliability simulation integrated in the SPICE simulator (Parthasarathy 2006) as opposed to a script written around the SPICE engine (Tu et al. 1993)), the reliability simulation of analog circuits is always limited to rather small circuits (typically less than 200 transistors). Simulation methods dedicated for digital circuit simulation are able to handle bigger circuits. In Wang et al. (2010), for example, the authors only calculate the signal activity using an RTL simulator, which significantly speeds up simulation time. However, this method is only applicable to digital standard cell designs, while the analog or RF part of a chip could be much more sensitive to transistor aging.

In conclusion, the reliability simulators discussed above are still far from perfect. The ideal reliability simulator, for analog circuits, should have the following properties:

- Support all important aging effects: HCI, NBTI, PBTI and TDDB.
- Support stochastic aging effects in sub-45 nm CMOS technologies: e.g. BTI in ultra-scaled CMOS (also see Sect. 3.3.4).
- Be capable of analyzing the combined impact of different transistor aging phenomena on important circuit performance parameters.
- Be capable of analyzing the correlation between process variations and circuit aging.
- Provide an extensive report on the impact of aging on the circuit including performance $= f(t)$, yield $= f(t)$, failure rate, detection of aging-sensitive spots in the circuit, generation of a degraded netlist, etc.
- Be capable of analyzing large circuits (>1000 transistors) in a reasonable time frame (i.e. in a few hours) with limited computational power (i.e. with a personal computer).
- Require minimal input from the circuit designer: he or she should not need profound knowledge on transistor aging or its impact on the performance of a circuit.
- Be capable of generating a portable model of a (sub)-circuit to evaluate the impact of process variations and circuit inputs on the circuit lifetime in a high-level environment such as MATLAB.
- Be compatible with existing SPICE simulators.

- Be compatible with all important device models such as BSIM3, BSIM4, PSP and EKV.

In the next chapter, a set of new reliability simulation methods will be presented. These methods are intended to solve some problems with the tools presented in this chapter and to have the properties listed above.

4.5 Conclusions

Transistor aging effects are a potential reliability problem for analog circuits processed in nanometer CMOS. To guarantee correct circuit operation throughout the desired lifetime of a circuit, circuit designers typically use large design margins. However, these margins significantly reduce the circuit performance and/or result in a large area and power overhead. Furthermore, reliable circuit operation is still not guaranteed. Circuit reliability simulators are therefore becoming a mandatory part of the circuit design flow when working with an advanced CMOS technology.

This chapter has given an overview of existing reliability simulators. Over time, various methods have been presented in literature, first to solve HCI problems and later to also estimate the impact of BTI effects. Some of these techniques have been adopted in commercial simulators to enable a designer to accurately simulate the impact of these effects on a circuit's operation. However, none of the reviewed simulators provides sufficiently accurate transistor aging models for the simulation of an analog circuit processed in a nanometer CMOS technology. Further, they are mostly limited to the simulation of rather small circuits and not able to properly analyze the impact of stochastic effects such as TDDB, BTI in sub-45 nm CMOS or the impact of process variations. The first problem (the availability of compact aging models) has been solved in Chap. 3, where accurate compact models for each important aging effect have been discussed. The simulation-related problems will be addressed next in Chap. 5.

Chapter 5
Analog IC Reliability Simulation

5.1 Introduction

In Chap. 4, an overview of existing reliability simulators has been given. Although a lot of research has been conducted in this area, leading to the implementation of a reliability simulation framework in each of the major commercial SPICE simulators, there are still a lot of deficiencies remaining (also see Sect. 4.4). Especially with the evolution to ever-smaller CMOS devices, statistical effects resulting from process variations and stochastic aging effects become more and more important. On top of that, most academic and commercial simulators are limited to the simulation of rather small circuits. Accurate reliability evaluation of large analog or mixed-signal circuits is therefore still not possible.

In this chapter a set of simulation methods, addressing these problems and other issues listed in Sect. 4.4, is proposed. The focus of the proposed simulator is on the simulation of analog circuits, although the methods can also be applied to small to medium-sized digital blocks. Section 5.2 first discusses an implementation of a deterministic reliability simulator. Such a simulator does evaluate the impact of transistor aging on the performance of the circuit, but does not include stochastic effects such as process variations or stochastic aging effects. The proposed method includes some techniques to achieve a good simulation accuracy while limiting the computational effort. The combined impact of multiple aging effects on a single transistor is also included. Further, a sensitivity analysis allows circuit weak spot detection and provides a designer with the necessary knowledge to design a more reliable circuit. Then, Sect. 5.3 discusses two implementations of a stochastic reliability simulator. This simulator includes the impact of stochastic effects and enables the capability to analyze the time-to-failure distribution of a design. A first implementation uses a brute-force Monte-Carlo approach which proves to be accurate but very computationally intensive. A second implementation using a response surface method is much more efficient. The latter also provides the user with an analytical model of the circuit performance as a function of the most important statistical parameters. The response surface method proves to be 1–3 orders of magnitude faster compared

E. Maricau and G. Gielen, *Analog IC Reliability in Nanometer CMOS*,
Analog Circuits and Signal Processing, DOI: 10.1007/978-1-4614-6163-0_5,

to the MC-based implementation. Finally, Sect. 5.4, proposes a flow to accurately simulate the impact of deterministic and stochastic aging effects on large analog and mixed-signal circuits. This method uses a hierarchical approach. First, the system is partitioned in system subblocks. Then, each subblock is modeled separately and finally the combined effect on the entire system is calculated using these subblock models. An innovative active learning sample selection strategy, combined with a fast function extraction symbolic regression, allows the fast and accurate modeling of each subblock. Each simulation method is first explained and then demonstrated on an example circuit. Section 5.5 the concludes the chapter.

5.2 Deterministic Reliability Simulation

A deterministic reliability simulator evaluates the impact of transistor aging effects on the performance of a circuit. Each aging effect is assumed to be deterministic: i.e. two identical transistors, stressed under the same conditions, will experience an identical aging-induced performance shift. Also, the influence of process variations is not considered. Although most of the existing simulators reviewed in Chap. 4 are deterministic, obtaining a good simulation accuracy, combined with short simulation times, is still a problem.

In this section, a state-of-the-art deterministic reliability simulator is proposed. First, the problem at hand is explained in more detail in Sect. 5.2.1. Then, Sect. 5.2.2 discusses the implementation details of the proposed reliability simulator. Finally, in Sect. 5.2.3, the simulator is demonstrated on an example circuit.

5.2.1 Problem Statement

A reliability simulator evaluates the behavior of a circuit as a function of the stress time T_{str}. This behavior is characterized with one or more performance parameters $\mathbf{P}^i$, such as the gain and bandwidth of an amplifier:

$$\mathbf{P}^i = \left[P_1^i, \ldots, P_m^i, \ldots, P_M^i\right] \tag{5.1}$$

with M the number of circuit performance parameters. t_i is the circuit age, with $i = \{0, \ldots, I\}$ and I the number of time points ($t_0 = 0\text{s}$ and $t_I = T_{\text{str}}$). Circuit aging is evaluated over a period of hours, months or even years ($T_{\text{str}} = 1\text{e}6\text{s} - 1\text{e}8\text{s}$). The period of an analog or a digital signal, however, is typically 15–20 orders of magnitude smaller ($T_{\text{sig}} = 1\text{e-}6\text{s} - 1\text{e-}12\text{s}$). Further, transistor aging is a nonlinear function of the applied voltages (also see Chaps. 2 and 3 where the aging effects and the corresponding models have been discussed). Therefore, when calculating circuit aging, it is not sufficient to use the average voltage applied to each transistor.

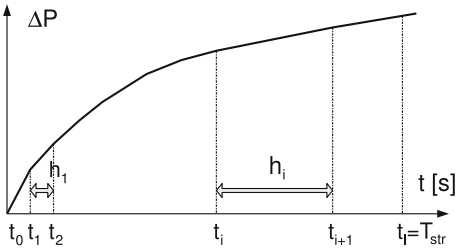


Fig. 5.1 To avoid simulation errors due to aging-induced bias voltage shifts, the reliability simulation is done in multiple steps. The time step between time point t_i and t_{i+1} is indicated as h_i. The total stress time is T_{str}

One has to include the actual time-varying stress present at each circuit node. To do this, most reliability simulators perform a transient simulation and calculate the aging of the circuit over a very small time span (i.e. a few signal periods). The result is then extrapolated to the desired stress time. However, besides the risk for extrapolation errors, the circuit bias conditions can change due to aging-induced transistor performance shifts. Therefore, the aging simulation is typically done in multiple steps, where each step h_i is only a fraction of the total stress time:

$$T_{\text{str}} = \sum_{i=0}^{I} h_i \tag{5.2}$$

where $i = 0$ corresponds to t = 0s and the number of steps I is determined by the required simulation accuracy. This concept is also depicted in Fig. 5.1. The circuit bias conditions change only marginally within one time step. Most simulators, reviewed in Chap. 4, use an approach where the number of time steps is fixed by the user. However, the required number of steps is very circuit dependent and is therefore often chosen too large (resulting in unnecessary long simulation times) or too small (resulting in inaccurate results). Further, within one time step, the aging is extrapolated from a simulation over a few signal periods to a few hours, days or even months. This can result in large extrapolation errors. Also, simulators such as BERT (see Sect. 4.2.1) calculate the impact of each aging effect separately, while in reality these aging effects affect each other. As a result, simulators such as BERT can significantly over- or underestimate the impact of transistor aging.

The next section proposes an improved deterministic reliability simulator which includes a set of methods to alleviate the problems described above.

5.2.2 Implementation

Most of the simulators in Chap. 4 are built around a SPICE engine. This approach is more flexible compared to a reliability simulator integrated in an existing SPICE engine. For that reason, the simulator discussed here, is also constructed as a script around an existing SPICE simulator. In this work, the Eldo SPICE simulator was used,

while the reliability simulation software was written in Python (Mentor graphics 2012; Python programming language 2012). First, the general simulation flow is explained. Then, algorithmic and implementation details are given and finally the speed and accuracy of the proposed method are discussed.

Simulation Flow

Figure 5.2 depicts a schematic representation of the deterministic reliability simulation flow. At the input of the simulator, the user has to provide a netlist, a stress bench and a test bench:

- The *netlist* describes the circuit under test. This circuit can consist of multiple subcircuits. Through additional statements in the netlist, the user can decide to let the entire circuit age or only a subset of transistors or subcircuits.
- The *stress bench* describes the input voltages and currents applied to the circuit during normal operation of the circuit. Also, it can be used to emulate an accelerated life test at a larger than nominal voltage and/or temperature. The latter is useful to determine suitable input conditions (i.e. test vectors) for prototype testing in a lab environment. The stress bench has to contain a transient analysis to extract accurate stress voltages on every node in the circuit.
- The *test bench* defines a number of tests done at regular time intervals during the simulated lifetime of the circuit. These tests are needed to measure the circuit performance parameters of interest. The latter can also be specified in the stress bench, in which case a test bench is not required.

In addition to the input files, the user must also specify three simulation parameters: the total stress time T_{str}, a minimum step size h_{min} and an upper limit for the simulation error ε_{max}. h_{min} and ε_{max} determine the simulation speed and accuracy and will be discussed in more detail further. The general flow of the reliability simulation algorithm itself is depicted in Fig. 5.2 and is as follows:

1. The input netlist is evaluated with the stress bench and the test bench, using a commercial SPICE simulator (indicated as *SPICE Simulation* in Fig. 5.2).
2. Circuit performance parameters are extracted and provided as an output to the simulator (indicated as *Performance PARAMETER Extraction* in Fig. 5.2). The time-dependent change in circuit performance is calculated in multiple time steps, corresponding to multiple runs through the simulation flow. During the first iteration of the algorithm, the performance values for a fresh circuit (at time t = 0s) are extracted. The simulation is stopped if the overall stress time exceeds T_{str}.
3. A step size algorithm is used to determine the next simulation step h_i (indicated as *STEP SIZE Algorithm* in Fig. 5.2). The minimum value for h_i is bounded by h_{min}, while ε_{max} determines the maximum step size.
4. For every transistor, the simulator calculates the total accumulated degradation during h_i (indicated as *GENERATE Aged Transistors* in Fig. 5.2). This calculation is done using the transistor aging compact models discussed in Chap. 3.

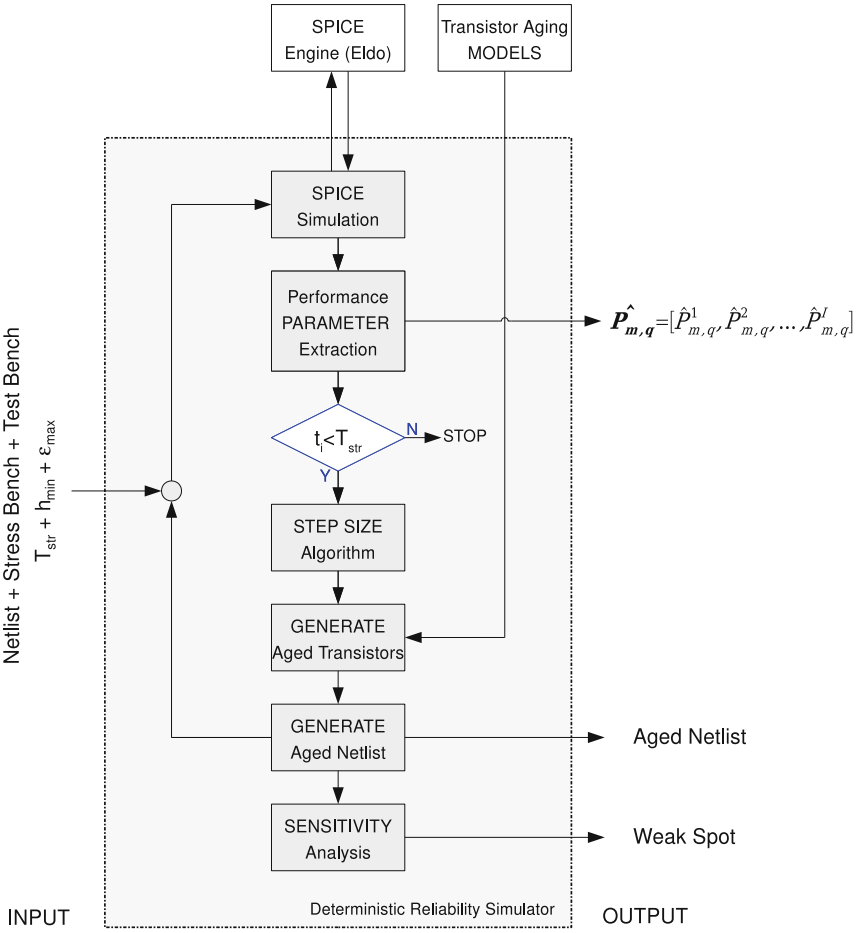


Fig. 5.2 Deterministic reliability simulation flow. The input to the simulator is a netlist with a stress bench, a test bench and a few simulation-related parameters. The output is an aged netlist at time T_{str} and a set of circuit performance parameters $\mathbf{P}^i$. The performance parameters are defined in the input netlist and evaluated at different time steps $t_i < T_{\text{str}}$. Also, weak spots are identified through the sensitivity analysis. These are transistors or subcircuits that cause the largest aging-induced circuit performance shifts

The voltages applied to each transistor, and provided as an input to the transistor aging models, are extracted from the SPICE simulation results obtained in the first step.

5. An aged netlist is generated (indicated as *GENERATE Aged Netlist* in Fig. 5.2). In this netlist, every transistor is replaced by an aging-equivalent transistor model as discussed in Sect. 3.5. This aged netlist is also provided as an output to the user.

6. The aged netlist is fed back to the SPICE simulator and the simulator calculates the circuit aging over the next time step h_{i+1}.
7. A sensitivity analysis, indicated as *SENSITIVITY Analysis* in Fig. 5.2, analyzes the sensitivity of the circuit output to the aging-induced changes in transistor parameters. To do this, the aged netlist is taken and in turn each transistor is replaced by its original (fresh) counterpart. Then, for each transistor the impact on the performance of the circuit, relative to the total aging-induced performance shift, is observed.

At the output of the simulator, the user receives a degraded version of the netlist for $t = T_{str}$. One can evaluate this aged netlist to compare the voltages and currents with the behavior of the original circuit. Or, alternatively, the degraded netlist can be used as part of a larger system, to see the impact of the aged circuit on the overall system performance. Further, the performance parameters, originally specified in the stress bench and the test bench, are provided at the output. These performance parameters have been evaluated after each time step and the user can use these to visualize the time-dependent circuit behavior. Finally, resulting from a sensitivity analysis on the aged netlist, circuit weak spots are pinpointed. A weak spot is a group of one or more transistors for which the aging has a large effect on the circuit performance.

In the next three sections, more details about the implementation of the reliability simulator are given. First, the simulator step size algorithm is discussed. Instead of a fixed or user-defined step size, the simulation steps are automatically determined such that the simulation speed is maximized while still maintaining good accuracy. Then, the algorithm to generate the aged transistors is reviewed. This algorithm allows to include the impact of multiple simultaneous aging effects. Also, it solves the simulation accuracy problem that typically arises when extrapolating from aging of a transistor over a few signal periods to transistor aging in one time step h_i. Finally, the sensitivity analysis used to detect circuit weak spots is discussed.

Adaptive Step Size Algorithm

As explained in Chap. 4, academic or commercial simulators either calculate the circuit aging in one step or they use a user-defined number of steps. These steps are then logarithmically or linearly spread over the stress time. Using multiple time steps does help to include the aging-induced shift of the circuit bias voltages. However, using a fixed number of steps and a fixed step size either results in excessive simulation times or inaccurate results.

In this work, the step size and the number of steps is automatically adapted to the change in circuit performance. If the circuit performance changes a lot, the time steps are chosen small in order to reduce extrapolation errors. If the circuit performance is very insensitive to circuit aging, however, the simulator can use large time steps to

speed up the reliability simulation. Thus, during every iteration i, the magnitude of the time step h_i is maximized. At the same time, the simulation accuracy is guaranteed by limiting the maximum relative shift of each circuit performance parameter P_m^i during h_i: i.e. $\left|\frac{P_m^{i-1}-P_m^i}{\min(P_m^{i-1},P_m^i)}\right| \leq \varepsilon_{\max}$ (see Algorithm 2). $\varepsilon_{\max}$ is a parameter chosen by the designer (typically $0.01 < \varepsilon_{\max} < 0.1$). After every iteration i, the aged circuit is fed back to the input of the transient simulator and resimulated to get an input for the next extrapolation (also see Fig. 5.2). This procedure is repeated until the total stress time equals the user-defined lifetime T_{str}.

Algorithm 2 Adaptive Step Size Algorithm

INPUT: $i, t_i, t_{i-1}, h_i, \mathbf{P}^{i-1}, \mathbf{P}^i$

if $i == 0$ **then**
  First iteration:
  $i = i + 1$
  $h = h_{\min}$
else
  Determine error vector $\vec{\varepsilon}$:
  $$\vec{\varepsilon} = \left|\frac{\mathbf{P}^{i-1} - \mathbf{P}^i}{\min(\mathbf{P}^{i-1}, \mathbf{P}^i)}\right|$$
  Calculate next time step:
  $$h = \max\left(h_{\min}, h_i \left[\frac{0.9\varepsilon_{\max}}{\max(\vec{\varepsilon})}\right]\right)$$
  Evaluate error vector $\vec{\varepsilon}$:
  if $\max(\vec{\varepsilon}) < \varepsilon_{\max}$ or $h_i == h_{\min}$ **then**
    Previous time step was good:
    $i = i + 1$
  else
    Previous time step was too large:
    Restore netlist to state t_{i-1}
    $i = i$
  end if
end if
Update stress time:
$t_i = t_{i-1} + h$
$h_i = h$

OUTPUT: i, t_i, h_i

Transistor Aging Algorithm

The reliability simulation, as schematically depicted in Fig. 5.2, calculates the circuit aging in multiple time steps to prevent extrapolation errors due to bias voltage shifts. For each time step h_i, a transient simulation is performed to extract the applied stress voltages on each transistor over a short stress time h_{tran}. Then, for each transistor, the accumulated stress during h_i is calculated. Yet, the former is much smaller than the latter:

$$h_{\text{tran}} \ll h_i \tag{5.3}$$

As discussed in Chap. 4, transistor aging is therefore typically simulated by first calculating the degradation over h_{tran} and then extrapolating the result to h_i. Unfortunately, this can result in large extrapolation errors.

Therefore, in this work an alternative simulation method is proposed. The stress waveform, obtained from the SPICE simulation, is stretched in time until it matches h_i (see Fig. 5.3). Then, when the accumulated aging-induced damage is calculated, there is no need for extrapolation since the damage at the end of the stress signal corresponds to the damage after stress time h_i. The major assumption behind this technique is that the aging effect is frequency independent. Indeed, by stretching the stress waveform

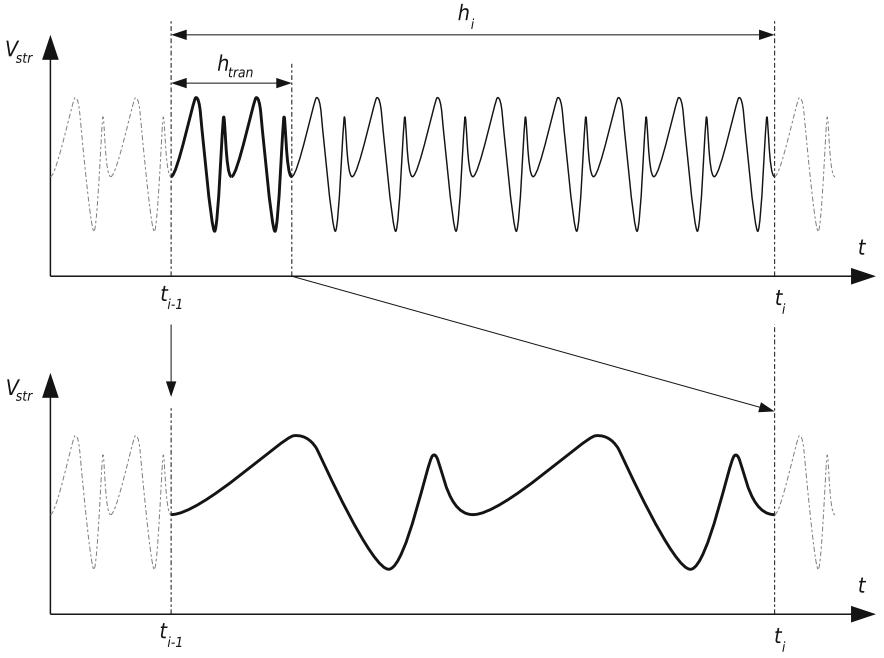


Fig. 5.3 Stretching the stress waveform to match the stress time reduces simulation time while at the same time avoiding the need for extrapolation

in time, a similar but much slower varying variant of the original stress waveform is used for the aging calculation. As already discussed in Chap. 2 and experimentally demonstrated by (Sasse 2008; Ramey et al. 2009), HCI and BTI effects are frequency independent.[1] Thus, the above method is valid and offers a better alternative to extrapolating the transistor aging effects and risking large simulation errors.

Multiple aging effects can simultaneously affect the behavior of a single transistor. For example, HCI and PBTI both affect the V_{TH} of an nMOS transistor. To simulate the combined impact of both effects, the aging of a single transistor is also performed in multiple time steps (see Algorithm 3). The algorithm uses an adaptive step size

Algorithm 3 Transistor Aging Algorithm

INPUT: t_i, h_i, D_{i-1}

$h' = h_i$

while $h_i > 0$ **do**

  Calculate aging over h' (also see chapter 3):
  $D' = D_{i-1} + D_{\mathrm{HCI}} + D_{\mathrm{BTI}}$

  if $D' - D_{i-1} < 0.1 D_{i-1}$ **then**
    Good step size:
    $h_i = h_i - h'$
    $D_{i-1} = D'$
  end if

  if $D' == D_{i-1}$ **then**
    Increase step size:
    $h' = \min(2h', h_i)$
  else
    Calculate new step size:

$$h' = \min\left(h_i, h' \frac{0.1}{\left|\frac{D' - D_{i-1}}{D_{i-1}}\right|}\right)$$

  end if

end while

Update transistor degradation:
$D_i = D'$

OUTPUT: D_i

algorithm, similar to the algorithm used for the overall reliability simulation (see Algorithm 2). Again, this approach maximizes the simulation speed, but at the same time limits extrapolation errors by limiting the maximum relative shift in transistor degradation.

[1] TDDB is frequency dependent (Sasse 2008). However, TDDB is also a stochastic effect and is therefore by default not supported by the kind of simulator presented in this section. A suitable approach, including TDDB, will be discussed in Sect. 5.3.

Sensitivity Analysis

After evaluating the circuit behavior as a function of the stress time, from a designer point of view, it is also interesting to know the cause of failure in case of an actual reliability problem. A sensitivity analysis enables a designer to identify possible weak spots in the circuit: i.e. transistors or subcircuits dominating the aging-induced circuit performance shift. The sensitivity of performance parameter P_j to the aging of transistor Mx is defined as (Sensitivity analysis 2012):

$$S^{P_m}_{\Delta P_{m,\mathrm{Mx}}} = \frac{\Delta P_{m,\mathrm{Mx}}}{\Delta P_m} \tag{5.4}$$

where ΔP_m represents the aging-induced performance shift when all transistors in the circuit are aging. $\Delta P_{m,\mathrm{Mx}}$ is the performance shift when all transistors in the circuit are aging, except for transistor Mx which does not age at all. The sensitivity value, given in Eq. (5.4), is evaluated for each parameter P_m and for each transistor Mx. If the value of $S^{P_m}_{\Delta P_{m,\mathrm{Mx}}}$ is large, the aging-induced shift of P_m is dominated by the aging of transistor Mx. The sensitivity value is positive when transistor aging results in an increase of P_m and negative if P_m decreases. Note how Eq. (5.4) does not only depend on the susceptibility of the circuit performance to transistor variations. Indeed, the aging-induced magnitude of these variations is also important. In other words, a transistor that ages a lot does not necessarily lead to circuit failure and vice versa.

Complexity and Accuracy

The simulation speed and accuracy of the proposed method are now validated on a current mirror test circuit. As circuit performance parameter P, the aging-induced shift in output current was monitored. The reliability simulation was done multiple times. First, a number of times with an increasing number of fixed time steps (in accordance with the method integrated in most aging simulators discussed in Chap. 4). Then, the same simulation was done with the adaptive step size algorithm for different values of ε_{max}. A smaller ε_{max} results in a smaller simulation error, but a larger number of time steps. For each simulation result, the normalized mean square error (NMSE) between the actual degradation[2] and the calculated degradation was calculated:

$$\mathrm{NMSE} = \frac{\sigma^2(\varepsilon_P)}{\sigma^2(P)} \tag{5.5}$$

[2] This value has been approximated, based on a simulation with a large number of fixed time steps (I = 1000).

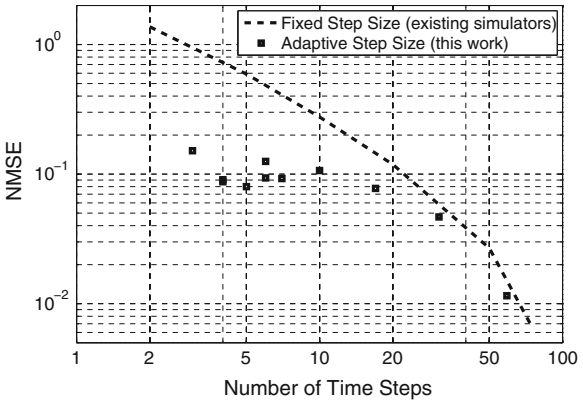


Fig. 5.4 Validation of the proposed reliability simulation flow. The prediction error (NMSE), when using the proposed algorithm with adaptive step size, is significantly lower compared to conventional methods using a fixed number of time steps

with $\sigma^2(\varepsilon_P)$ the variance on the prediction error and $\sigma^2(P)$ the variance on the stress-time-dependent output current. The test results are depicted in Fig. 5.4. The figure clearly shows how the proposed algorithm yields a much lower prediction NMSE compared to the conventional fixed step size approach. This difference is the largest (with an up to 10 times smaller prediction error for the proposed algorithm) for a small number of time steps. In reality, that is also the region of interest, since a small number of time steps results in short simulation times. The proposed algorithm therefore results in short reliability simulation times, combined with a good simulation accuracy.

5.2.3 Circuit Example

Above, the flow of the deterministic reliability simulator has been discussed in detail. In this section, the method is demonstrated on an clocked comparator circuit. First, the circuit topology and the applied stress voltages are discussed. Next, the first-order aging model, presented in Sect. 3.6, is used to calculate the impact of transistor aging on the circuit performance. This hand calculation is then compared to the more accurate result obtained from the reliability simulator. The output of the simulator is also discussed in detail.

Circuit Schematic

The example circuit is a fully differential clocked comparator as depicted in Fig. 5.5. The circuit is simulated in a 32 nm CMOS predictive technology with a 1 V supply

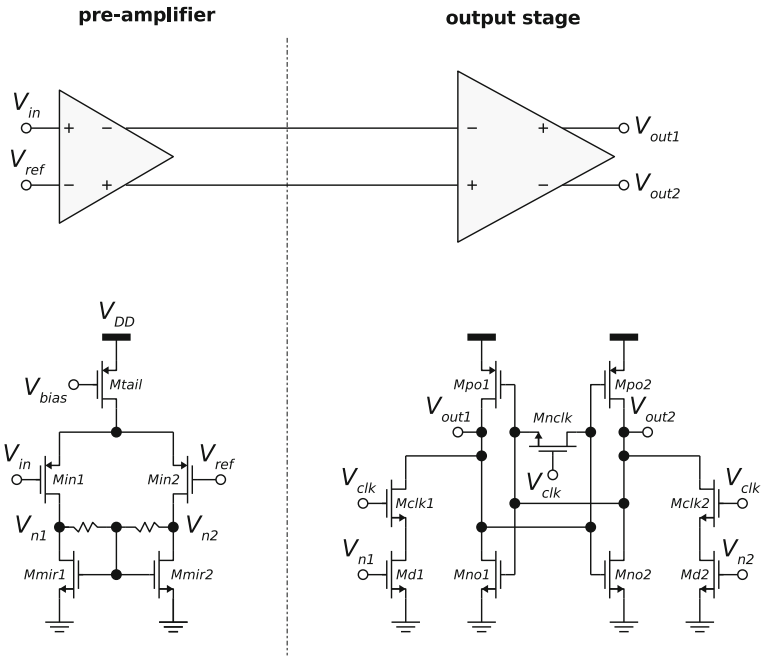


Fig. 5.5 The clocked comparator, used as a demonstrator circuit for the deterministic reliability simulator

voltage (Arizona state university predictive technology model 2012). At one input a reference voltage V_{ref} is applied, which is compared to the voltage applied at the other input V_{in}. If $V_{\text{in}} > V_{\text{ref}}$, V_{out1} is high and V_{out2} is low. In a first stage, the input is pre-amplified and then, when the clock signal is low, it is converted into a digital output by a back-to-back inverter in the output stage.

In this example, $V_{\text{ref}} = 0.2\,\text{V}$ and V_{in} is a sine wave with an amplitude of 0.4 V and a DC value of 0.5 V (see Fig. 5.6). If $V_{\text{in}} > V_{\text{ref}}$, the stress voltage on the gate of transistor Min2 is larger than the voltage on the gate of Min1. Also, the stress on Md2 will be larger than the stress on Md1. As a consequence, Min2 and Md2, which age due to NBTI and PBTI respectively, will degrade more than Min1 and Md1 (as indicated in Fig. 5.6). Next, if the input voltage is low and $V_{\text{in}} < V_{\text{ref}}$, the situation is reversed and Min1 and Md1 will age more than Min2 and Md2. However, in this example, the latter only happens during a small fraction of the time. Overall, due to this asymmetric stress applied at the circuit input, Min2 and Md2 will age more than Min1 and Md1. This results in a time-dependent increase of the mismatch between the transistors in the circuit and in turn affects the circuit performance. Here, two performance parameters are observed: the input offset and the slew rate.

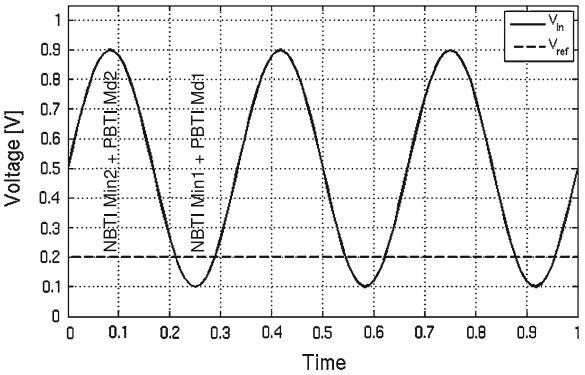


Fig. 5.6 Input waveform applied to the example circuit. When $V_{\text{in}} > V_{\text{ref}}$, Min2 and Mo2 suffer from NBTI and PBTI stress respectively. When $V_{\text{in}} < V_{\text{ref}}$ the situation is reversed, but over a shorter time period

Hand Calculations

In Sect. 3.6 a first-order transistor aging model, intended for hand calculations, has been proposed. Here, the use of this model is demonstrated by calculating the aging-induced shift in offset and slew rate for the example circuit.

To first order, the circuit offset is determined by the mismatch between the input pair of the first and second stage:

$$V_{\text{offset}} \approx \delta V_{\text{TH,Min}} + \frac{\delta V_{\text{TH,Md}}}{A_{\text{pre}}} \tag{5.6}$$

with $\delta V_{\text{TH,Min}}$ and $\delta V_{\text{TH,Md}}$ the mismatch between transistor Min1 and Min2, and transistor Md1 and Md2, respectively. A_{pre} is the gain of the pre-amplifier, which is in this case 12dB. The stress voltage applied at the input V_{in} of the comparator is time-varying (see Fig. 5.6). The first-order aging model, however, does only support DC stress voltages. Therefore, for this calculation, the average DC stress of 0.5 V is used as the stress voltage at the circuit input. A SPICE DC operating point analysis, with 0.5 V at the circuit input, yields the stress voltages applied to each transistor in the circuit and provides the necessary input for the reliability hand calculations. The result of these calculations, for the transistors of interest and for a stress time of 5 years, are depicted in Table 5.1. Transistors Min1 and Md1 do not degrade, therefore Eq. (5.6) can be simplified to:

$$\begin{aligned} V_{\text{offset}} &\approx |\Delta V_{\text{TH,Min2}}| + \frac{|\Delta V_{\text{TH,Md2}}|}{A_{\text{pre}}} \\ &= 4.6\,\text{mV@5year} \end{aligned} \tag{5.7}$$

Table 5.1 Results of the hand calculation for the aging of the most important transistors in the amplifier, evaluated with the model in Sect. 3.6

	$V_{GST}(V)$	$\lvert\Delta V_{TH}\rvert$ (mV)@5 year
Min1	−0.12	0.0
Min2	0.17	1.9
Md1	−0.02	0.0
Md2	0.31	10.7
Mtail	0.21	3.2

The anticipated first-order offset after 5 years of operation is therefore 4.6 mV. The circuit slew rate can be approximated by:

$$\mathrm{SR} \approx \frac{I_{\mathrm{SD,Mtail}}}{C_{\mathrm{L}}} \tag{5.8}$$

with $I_{\mathrm{SD,Mtail}}$ the drain-source current of Mtail. C_{L} is the load capacitance at the output of the circuit. Only $I_{\mathrm{SD,Mtail}}$ can change due to transistor aging. Table 5.1 shows how the expected V_{TH} shift for Mtail is 3.2 mV. The overdrive voltage of that transistor, however, is 210 mV. Therefore, the aging-induced shift of the tail current and the slew rate is expected to be very small.

As demonstrated above, the hand calculations can help to obtain a first impression of the impact of transistor aging on the performance of a circuit. However, the aging model used here is only a first-order model which takes the average voltage applied to each transistor as stress voltage. This calculation therefore only indicates a potential reliability problem and a computer reliability simulation is required to obtain a more accurate result.

Simulation Results

The aging-induced shift in circuit performance has also been calculated by the simulator discussed above. The results are depicted in Fig. 5.7. The input offset changes from nearly 0–1.27 mV over a time span of five years. The slew rate, which mostly depends on the transistor drive current and the output capacitance, remains fairly constant. The change in both performance parameters has a log(t) time dependence, which corresponds to the logarithmic time dependence of the NBTI and PBTI effect (also see Chap. 3). Compared to the offset shift of 4.6 mV calculated by hand in the previous section, the actual offset shift of 1.27 mV is much smaller. This is because, in contradiction to the assumed DC input when calculating the offset by hand, the actual input is a sine wave. Therefore, transistors Min1 and Md1 will also age, although not as much as Min2 and Md2. As expected from the hand calculations, the shift in slew rate is indeed very small. The results obtained in the previous section therefore show to be very useful to get a first-order estimate for the circuit aging. However, an actual reliability simulation is required in order to get a more accurate result. Further, the reliability simulation also returns additional results to

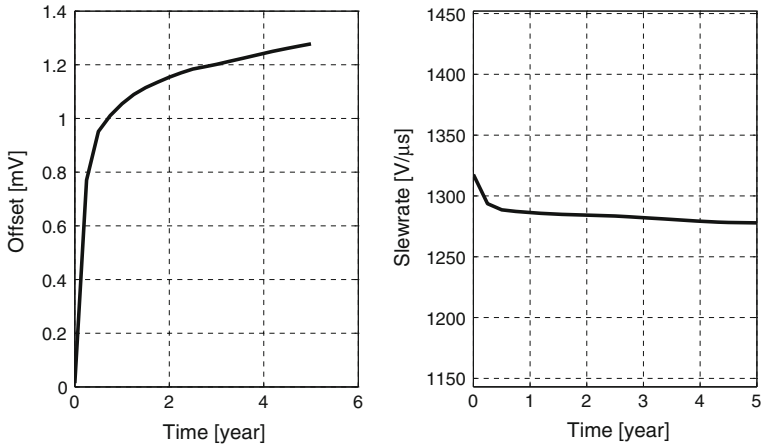


Fig. 5.7 The aging-induced circuit performance shift. The input offset is very sensitive to circuit aging, while the slew rate remains more or less constant

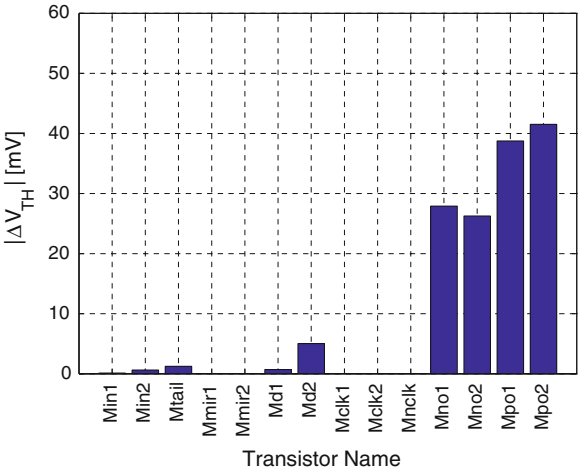


Fig. 5.8 The aging-induced $|V_{TH}|$ shift of every transistor in the circuit

allow an even better understanding of the impact of transistor aging on the circuit performance.

Figure 5.8 shows the $|\Delta V_{TH}|$ for every transistor in the circuit. The V_{TH} shift of the output stage transistors is very large, compared to the shift in the other transistors, although still rather small in an absolute sense ($\max(\Delta V_{TH}) = 41\,\text{mV}$). The reason for these relatively large V_{TH} shifts is the application of large stress voltages to the transistors in the output stage ($V_{GS} = V_{DD}$ when switched on). Still, as explained above, the aging of the output transistors does not have a large impact

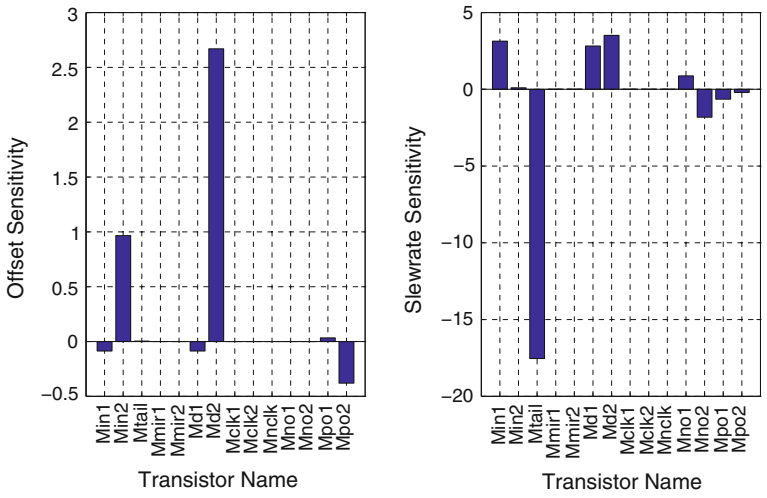


Fig. 5.9 A sensitivity analysis of the impact of transistor aging on the performance of the circuit reveals the transistors that are responsible for potential circuit reliability problems

on the performance the circuit. As such, Fig. 5.8 only provides limited information about the degradation of the circuit itself. The results of the sensitivity analysis, however, provide more relevant information and are depicted in Fig. 5.9. Figure 5.9 (left) shows how the offset shift mostly results from degradation in transistors Min2 and Md2. This corresponds to the intuitive analysis and the results of the hand calculation discussed above. The degradation of transistor Mtail, which has the largest impact on the drive current, has the biggest impact on the slew rate (see Fig. 5.9 (right)). Note how the sensitivity analysis only compares the relative contributions of the different transistors in the circuit. One still needs to examine the overall circuit degradation (see Fig. 5.7) to assess whether there is an actual reliability problem or not.

Chouard et al. (2010b, 2011) conducted stress measurements on differential amplifier circuits processed in a 32 nm CMOS technology. The results of these experiments are in accordance with the simulations presented here. A deterministic reliability simulator, as discussed in this section, is therefore a useful tool for a designer to assess the impact of device aging on his or her circuit. Nevertheless, this simulator does only include deterministic aging effects. The impact of stochastic effects, such as TDDB, process variations or even BTI and HCI effects in ultra-scaled CMOS processes cannot be simulated. Simulators published in literature or commercially available also do not support this (also see Chap. 4). In the next section, this problem is studied in more detail and a solution is proposed.

5.3 Stochastic Reliability Simulation

Section 5.2 has discussed the implementation of a deterministic circuit reliability simulator. Such a simulator calculates the effect of transistor aging on the performance of a circuit, but does not include the impact of process variations and stochastic aging effects such as RDF, LER and TDDB. In nm CMOS technologies, however, parametric process variability can have a significant impact on the performance of a circuit right after production, but also when it ages (Bestory et al. 2007; Asenov et al. 2008; Huard et al. 2008). No circuit sample is exactly the same, therefore each circuit ages differently and some circuits will fail earlier than others. This dispersion of the circuit time to failure (TTF) is also affected by stochastic aging effects such as TDDB and BTI and HCI in sub-45 nm CMOS.

In this section, a stochastic circuit reliability simulator is proposed. Such a simulator includes the impact of stochastic effects and enables a designer to calculate the time-to-failure distribution, rather than only the mean time to failure. Section 5.3.1 first explains the problem in more detail. Then, in Sect. 5.3.2, a first implementation of a stochastic reliability simulator is proposed. This implementation uses a Monte-Carlo approach, which is accurate but requires long simulation times. A more efficient approach, using design of experiments to develop a circuit model, is proposed in Sect. 5.3.3. Finally, Sect. 5.3.4 demonstrates both implementations of the stochastic reliability simulator on an example circuit.

5.3.1 Problem Statement

In Chap. 2, two separate unreliability effects were identified: spatial unreliability effects, which are fixed in time, and temporal unreliability effects, which are time-dependent. Each of these effects can be considered to be deterministic or stochastic.

Process variations are spatial stochastic aging effects. Due to these effects, matched transistors are, in reality, not completely identical. As a consequence they will not age in exactly the same way. Assume for example a pMOS transistor which ages due to the NBTI effect. If a fixed stress voltage is applied to the gate of the transistor, the aging-induced change in V_{TH} can be modeled as[3]:

$$V_{TH} = V_{TH0} + \underbrace{(V_{GS} - V_{TH0})^{\alpha}(C + n\log(t))}_{\Delta V_{TH}} \tag{5.9}$$

[3] This is a first-order approximation to illustrate the problem. A more complete and accurate transistor compact model has been discussed in Chap. 3.

where V_{TH0} is the initial transistor threshold voltage. α, C and n are technology-dependent parameters. The impact of process variations can be modeled as variation on V_{TH0}[4]:

$$V_{\mathrm{TH0}} = \mathcal{N}\left(\mu(V_{\mathrm{TH0}}), \sigma(V_{\mathrm{TH0}})\right) \tag{5.10}$$

From Eqs. (5.9) and (5.10) one can now derive a first-order expression for the standard deviation on V_{TH}, as a function of transistor age:

$$\sigma(V_{TH}) = \sigma(V_{\mathrm{TH0}})\left[1 - \alpha(V_{\mathrm{GS}} - V_{\mathrm{TH0}})^{\alpha-1}(C + n\log(t))\right] \tag{5.11}$$

$$= \sigma(V_{\mathrm{TH0}})\left[1 - \frac{\alpha \Delta V_{\mathrm{TH}}}{V_{\mathrm{GS}} - V_{\mathrm{TH0}}}\right] \tag{5.12}$$

Equation (5.11) suggests a relationship between the initial transistor variability and the transistor age. In other words, the initial mismatch between transistors in a circuit changes over time when these transistors age, even if they are subjected to the same stress voltages. Further, temporal stochastic unreliability effects such as TDDB will result in additional time-dependent circuit variability. In older technologies (>65 nm CMOS), BTI and HCI can be considered as temporal deterministic unreliability effects. However, for circuits processed in sub-45 nm CMOS technologies even BTI and HCI can no longer be approximated as a deterministic phenomenon (also see Chap. 3). This also corresponds with measurements reported in (Huard et al. 2008), which show an increase of the V_{TH} standard deviation over time.

Spatial stochastic reliability effects can have a significant impact on the performance of a circuit. The pool of potentially fabricated circuits resulting from these variations, can be described by an N_{s}-dimensional circuit factor space $\mathcal{F}$, where every dimension or factor $f_{n_{\mathrm{s}}}$, $n_{\mathrm{s}} = \{1, \ldots, N_{\mathrm{s}}\}$, represents a technology parameter following a process-dependent statistical distribution. Examples of these factors are a resistor value, the transistor V_{TH} and the transistor current factor β. Every factor $f_{n_{\mathrm{s}}}$ can be characterized by a mean and a standard deviation. Data for each of these factors can come from test structures on wafers or from SPICE simulations on statistical models that have been characterized by the foundry. Additionally, temporal stochastic reliability effects, such as TDDB or BTI in sub-45 nm CMOS, also have an impact on the failure distribution. Therefore, $\mathcal{F}$ is augmented with N_t temporal factors $f_{n_{\mathrm{t}}}$, $n_{\mathrm{t}} = \{1, \ldots, N_{\mathrm{t}}\}$. Examples of these temporal factors are parameters β and θ in the transistor breakdown model (see Sect. 3.5). The total number of factors in $\mathcal{F}$ is therefore $N = N_{\mathrm{s}} + N_{\mathrm{t}}$. One point in $\mathcal{F}$ corresponds to one statistical circuit sample and is represented as $\mathbf{f}_q = [f_{q,1}, \ldots, f_{q,N}]$. An example two-dimensional factor space is shown in Fig. 5.10 (upper left) with factors f_1 and f_2. One circuit sample $\mathbf{f}_q$ is also indicated.

[4] For simplicity of explanation, the process-induced threshold voltage variation is assumed to follow a Gaussian distribution. In reality, this is not necessarily the case. The conclusions made in this section are however also valid for other distributions.

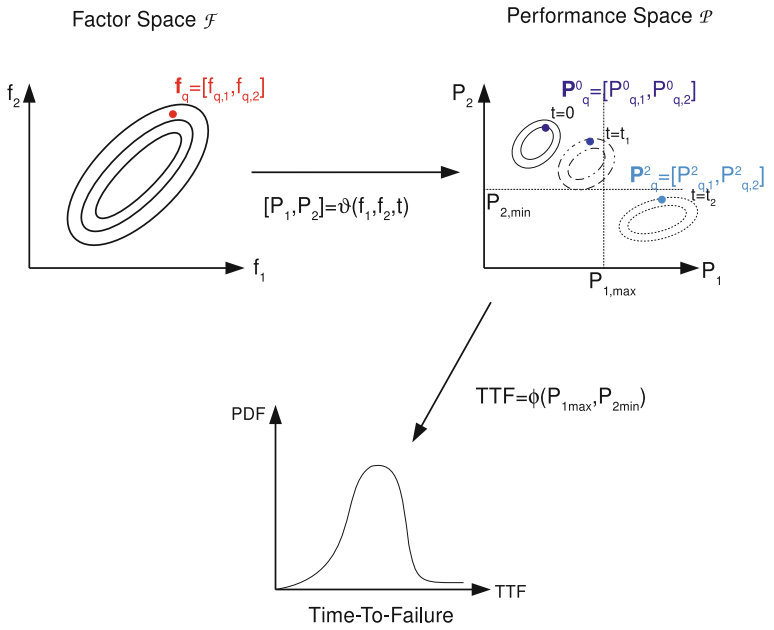


Fig. 5.10 Schematic representation of a 2-dimensional factor space $\mathcal{F}$ (*upper left*), resulting in a 2-dimensional performance space $\mathcal{P}$ (*upper right*). Every point in $\mathcal{F}$ maps on a point in $\mathcal{P}$. Due to temporal deterministic and stochastic reliability effects the latter shifts over time, while the spread on all points in $\mathcal{P}$ also changes due to spatial and temporal stochastic reliability effects. Adding circuit specifications ($P_{1,\max}$ and $P_{2,\min}$) results in a time-to-failure distribution TTF (*bottom*)

$\mathbf{f}_q \in \mathcal{F}$ has a behavior defined by M circuit performance parameters (e.g. DC gain, bandwidth, offset, etc.), corresponding to a point in an M-dimensional circuit performance space $\mathcal{P}$. At time $t_0 = 0$s, every circuit sample $\mathbf{f}_q \in \mathcal{F}$, can be mapped on a point $\mathbf{P}^0 = [P^0_1, \ldots, P^0_M] \in \mathcal{P}$. In Fig. 5.10 (upper right) two performance parameters P_1 and P_2 are shown. When the circuit ages, the behavior of the circuit changes. At time t_i, this will result in a new point $\mathbf{P}^i$. On Fig. 5.10, the initial performance of sample $\mathbf{f}_q$ is indicated as $\mathbf{P}^0_q$; at time t_2 this changes to $\mathbf{P}^2_q$. Due to the spatial and temporal unreliability effects, both the average and the spread of each performance parameter will change. The link between $\mathcal{F}$ and $\mathcal{P}$ is defined as ϑ:

$$\mathbf{P}^i_q = \vartheta(\mathbf{f}_q, t_i) \text{ with } \mathbf{f}_q \in \mathcal{F} \text{ and } \mathbf{P}^i_q \in \mathcal{P} \tag{5.13}$$

When integrated as part of a larger system or product, the circuit has to meet certain application-dependent specifications. These specifications are indicated as $\mathbf{P}_{\min}$ and $\mathbf{P}_{\max}$ in Fig. 5.10 (upper right). Every circuit violating these specifications is considered a failure and, since every circuit ages differently, this results in a time-to-failure distribution (see Fig. 5.10 (bottom)):

$$\text{TTF} = \phi(\mathbf{P}_{\min}, \mathbf{P}_{\max}) \tag{5.14}$$

A stochastic reliability analysis tool, as discussed in this section, quantifies the link between the factor space, the performance space and the TTF through the functions ϑ and ϕ.

5.3.2 Implementation 1: Monte-Carlo Simulation

Above, the necessity to include the impact of stochastic unreliability effects in a reliability simulator has been explained. In this work, two implementations of such a stochastic reliability simulator are explored. A first approach, discussed in this section, uses Monte-Carlo (MC) simulations to emulate the manufacturing process by simulating a number of randomly selected samples $\mathbf{f}_q \in \mathcal{F}_{Q_{\text{MC}}}$, with $\mathcal{F}_{Q_{\text{MC}}} \subset \mathcal{F}$:

$$\mathbf{f}_q,\ q = \{1, \ldots, Q_{\text{MC}}\} \tag{5.15}$$

with Q_{MC} the number of Monte-Carlo samples. Below, the simulation flow and accuracy are discussed.

Simulation Flow

The simulation flow is depicted in Figs. 5.11 and 5.12. At the input of the simulator, the user must provide a netlist with corresponding stress bench and test bench. In addition, the user can also set circuit specifications defining the minimum and the maximum values for each circuit performance parameter (e.g. the minimum gain and the maximum offset of an amplifier). The simulation flow itself is as follows:

1. The circuit factors are extracted from the input netlist (indicated as *Circuit Factor EXTRACTION* in Figs. 5.11 and 5.12). These factors determine the circuit factor space $\mathcal{F}$.
2. A set of random samples $\mathcal{F}_{Q_{\text{MC}}}$ is selected from the factor space (indicated as *MC Sample SELECTION* in Figs. 5.11 and 5.12). The number of MC samples is determined by the designer and is typically between 100 and 500. Further, the samples are selected according to their probability of occurence. This probability is a function of the distribution of each factor in $\mathcal{F}$ and depends on the technology. In this work, spatial factors and their corresponding distributions were extracted from data provided by the foundry: i.e. using Monte-Carlo transistor models. Variations on temporal factors were extracted from literature and measurements (also see Chap. 3).
3. For each sample, $\mathbf{f}_q \in \mathcal{F}_{Q_{\text{MC}}}$ corresponding to a unique combination of spatial and temporal factor values, the aging-induced degradation is calculated (indicated as *Deterministic Reliability SIMULATION* in Fig. 5.11). The parameter variations

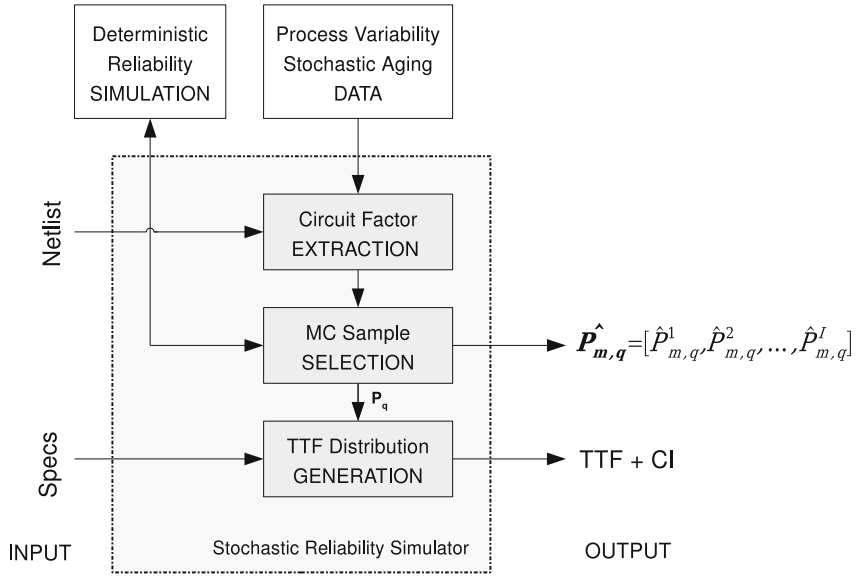


Fig. 5.11 Simulation flow of the Monte-Carlo-based implementation of the stochastic reliability simulator

and aging-induced degradation of one sample are deterministic. Therefore, each sample can be simulated with a deterministic reliability simulator as discussed in Sect. 5.2. To reduce simulation time, different samples can be calculated in parallel (on multiple computer nodes). The deterministic reliability simulator returns the time-dependent performance shift $\mathbf{P}_{m,q}$ for each circuit sample $\mathbf{f}_q$.

4. If circuit specifications are provided, the TTF distribution is calculated (indicated as *TTF Distribution GENERATION* in Figs. 5.11 and 5.12). For every time point t_i, the simulator evaluates whether each sample $\mathbf{f}_q$ meets the specifications. Samples that do not meet all the specifications are labeled as failures. Eventually, the number of failures at each time point determines the TTF distribution. Further, bootstrapping is used to determine the confidence intervals on the TTF values (Efron 1979; Davison and Hinkley 1997). The bootstrapping method works as follows. First, the pool of MC-simulated circuit samples is randomly resampled for a large (>100) number of times. Then, for each of the resampled data sets, the TTF is generated and the properties of the distribution are calculated (e.g. the mean μ(TTF) and the standard deviation σ(TTF)). Finally, the spread on these property values is a measure for the uncertainty on that property. This method does not assume any underlying TTF distribution function and is valid as long as the original set of MC-samples is representative for the total population.

At the output of the simulator, the user obtains a set of points in $\mathcal{P}$, which shift as a function of the stress time. If circuit specifications are given, an estimation for the circuit time-to-failure distribution is also provided.

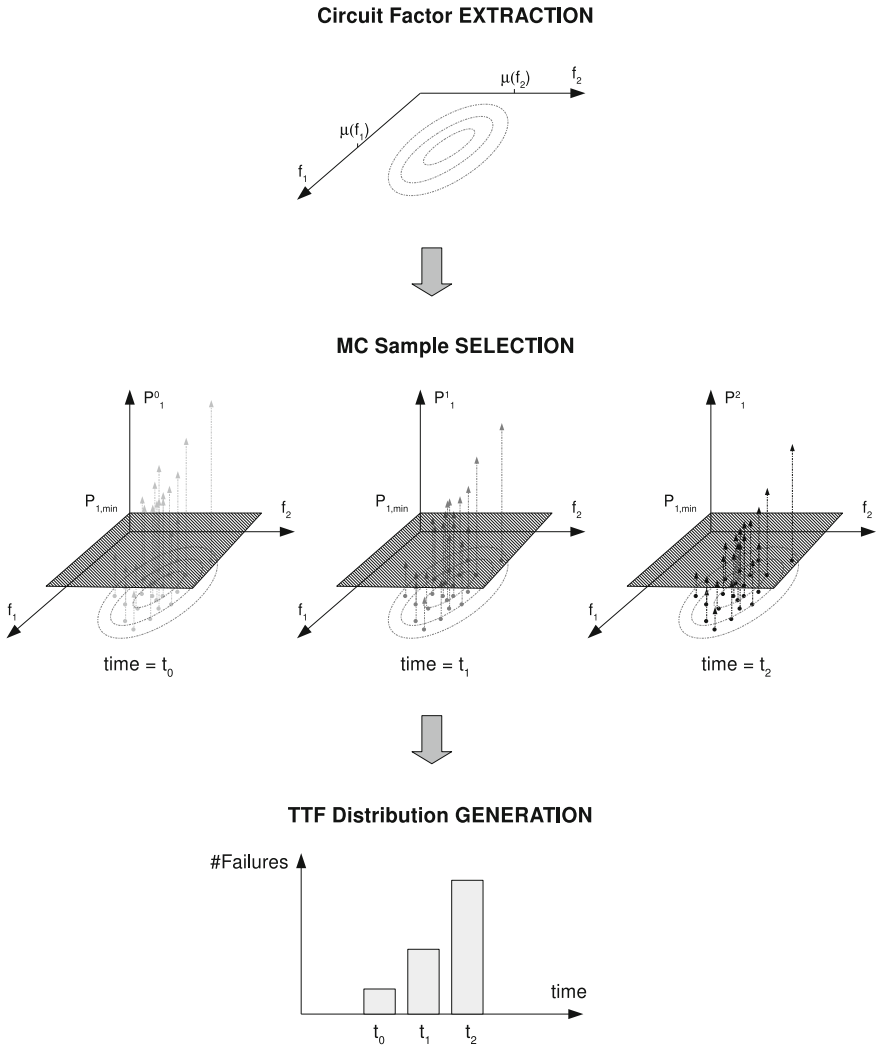


Fig. 5.12 Schematic representation of the Monte-Carlo-based implementation of the stochastic reliability simulator, applied to an example circuit with two input factors and one performance parameter

Complexity and Accuracy

A Monte-Carlo simulation is straightforward and accurate. The only assumption behind the method is that the set of random samples $\mathcal{F}_{Q_{\mathrm{MC}}}$ is representative for the total population $\mathcal{F}$. The error on the mean of the distribution at the output (in this case the standard error on the mean time to failure MTTF, $E_{\mu(\mathrm{TTF})}$) decreases with the square root of the number of Monte-Carlo samples:

$$E_{\mu(\mathrm{TTF})} \propto \frac{1}{\sqrt{Q_{\mathrm{MC}}}} \tag{5.16}$$

Further, if a sufficient number of samples are simulated, simulation errors will be dominated by the accuracy of the transistor aging compact models and the accuracy of the deterministic reliability simulator. The latter depends on the circuit under test and the settings of the deterministic reliability simulator. Although accurate, this MC-based method is very inefficient. The vast majority of samples will be concentrated near the nominal design and will provide very little extra information about $\mathcal{P}$ and the TTF distribution. Further, each sample is evaluated with a deterministic reliability simulation which is computationally very intensive, since it requires multiple transient simulations (see Sect. 5.2). Therefore, a more efficient method, requiring less reliability simulations, is required. Such a method is proposed in the next section.

5.3.3 Implementation 2: A Response Surface Methodology

A MC-based approach, as explained in Sect. 5.3.2, is accurate and easy to implement. However, the method is inefficient and computationally very intensive. Even when parallel computation is applied, this results in long overall simulation times. Furthermore, in case a reliability problem is detected (e.g. if 20 % of the circuits fail before the intended product lifetime), a Monte-Carlo simulation does not provide information on how to improve the circuit such that the problem can be reduced or solved. Therefore, a more efficient solution is required.

As discussed in Sect. 5.3.1, the objective of a stochastic reliability simulator is to quantize the relationship ϑ between the factor space and the performance space. A MC-based approach does this by simulating a finite amount of individual samples in $\mathcal{F}$. However, evaluating one combination of factors in $\mathcal{F}$, corresponding to $\vartheta(\mathbf{f}_q)$, is done with a deterministic reliability simulation and requires a large computational effort. A mathematical circuit model $\hat{\vartheta}$, however, requires much less evaluation time. Therefore, such a model allows very fast yield calculations (i.e. Monte-Carlo simulations on a model). Further, this model relates the circuit factors to the performance parameters and can therefore be used to optimize the circuit reliability and to localize weak spots in case of a reliability problem. The key issue is to find $\hat{\vartheta}$ such that it is accurate but also requires very few deterministic reliability simulations to be built.

In this work, design of experiments (DoE) techniques are used to build an accurate circuit model $\hat{\vartheta}$. In general, DoE or experimental designs are techniques to gather information about an unknown system where variation is present (Montgomery 2008; Engineering statistics handbook 2012). This variation is not necessarily under full control of the experimenter. Typically, the experimenter is interested in the relationship between some of the explanatory variables (system input) and the observed response variables (system output). DoE is about setting up an efficient and systematic procedure for doing the experiments such that analysis of the obtained data yields

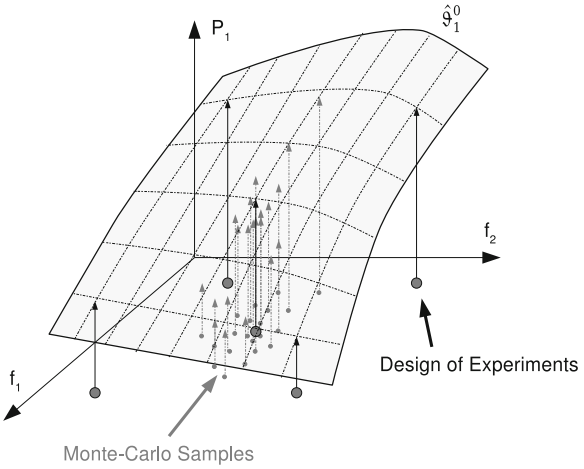


Fig. 5.13 A schematic representation of a two-dimensional factor space (defined by f_1 and f_2), projected on a one-dimensional performance space (defined by P_1). A Monte-Carlo simulation evaluates random samples in the factor space and requires a lot of simulations to obtain an accurate impression of the performance space. A DoE-based approach only simulates specific samples and then builds an analytical model $\hat{\vartheta}_1^0$ of the performance space. This model requires very little evaluation time

valid and objective conclusions. In addition, all of this is done under the constraint of minimizing the amount of experiments to limit run time and cost. There are four general problem areas for which DoE can be applied (Engineering statistics handbook 2012):

1. To *compare* the impact of different inputs on the behavior of the system.
2. To *screen* the system. In this case the experimenter wants to know which system inputs have the largest impact on the system output and which inputs are unimportant.
3. To *model* the system behavior with a mathematical model of the observed system outputs as a function of the controllable system inputs.
4. To *optimize* the system output.

In this work, the objective is to find an accurate model for the unknown function ϑ. The DoE determines which combinations of factors in $\mathcal{F}$ need to be simulated in order to develop this model. The simulation results can then be used to create $\hat{\vartheta}$. This concept is also illustrated in Fig. 5.13. MC-samples are typically taken near the nominal design point and therefore require a lot of simulations to obtain an accurate result. A few well chosen experimental design points, however, are already sufficient to build a circuit model. In literature, techniques like this have been proposed in the context of circuit yield simulation and optimization (Elias 1994; Leary 1995; Jing et al. 2004). Here, the DoE is optimized for reliability analysis. Also, the analysis is divided into two steps. First, a screening design is used to find the dominant input

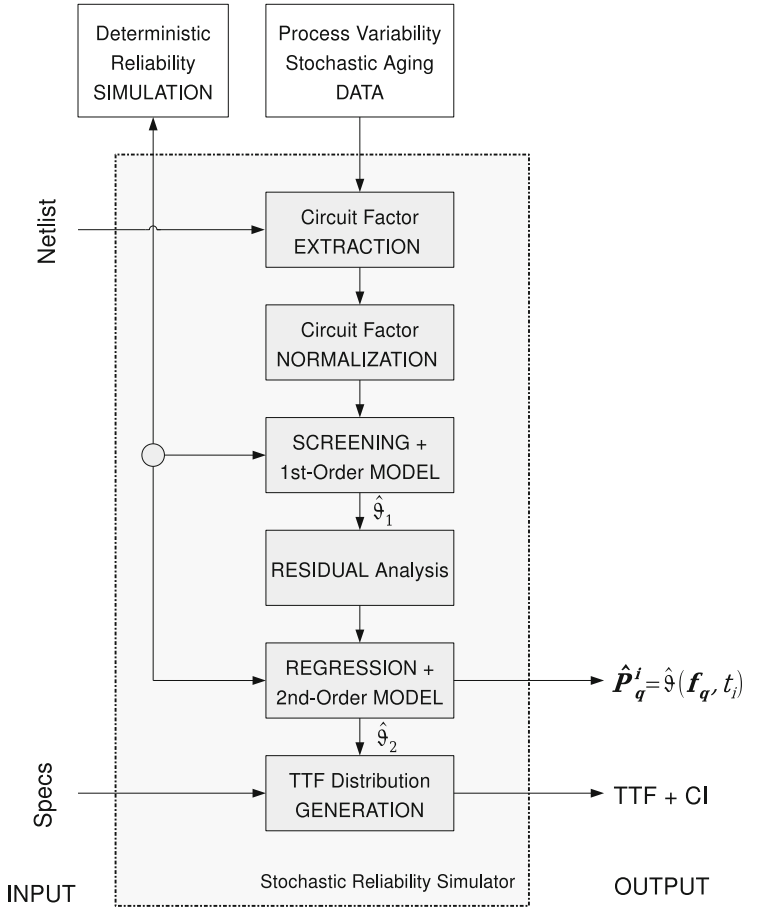


Fig. 5.14 A stochastic reliability simulation flow using design of experiments and a regression model to characterize the relationship between the factor space and the performance space

factors and to develop a first-order model. Then, a regression design is used to further refine the model where needed. The proposed method is explained in detail in the following sections.

Simulation Flow

Figures 5.14 and 5.15 illustrate the simulation flow for the DoE-based stochastic reliability simulator. The input to the simulator is a fresh (i.e. unstressed) netlist, a stress bench and a test bench. Also, a few simulation parameters such as the stress time T_{str} are required. The simulation flow itself is as follows:

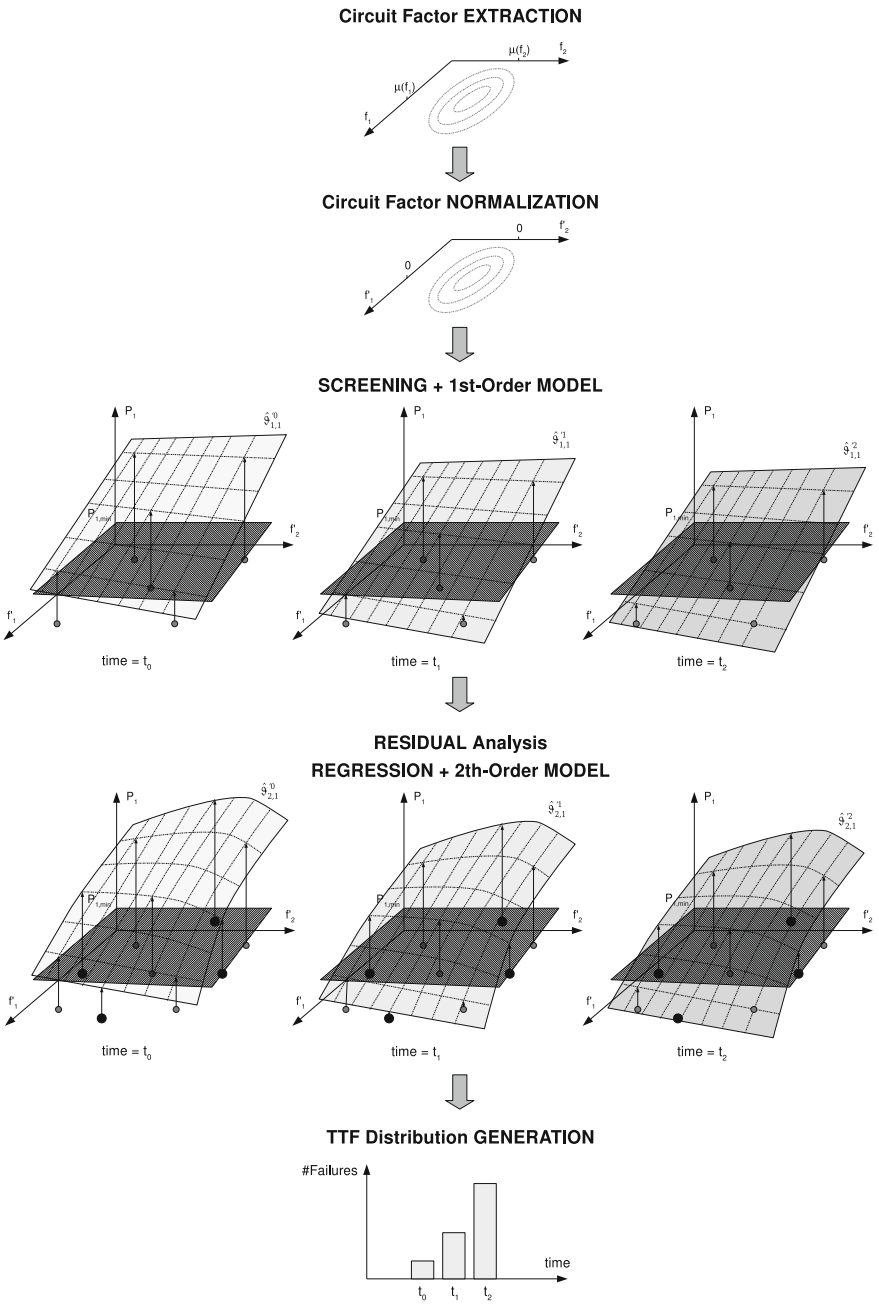


Fig. 5.15 A stochastic reliability simulator using design of experiments and a regression model to characterize the relationship between the factor space and the performance space, demonstrated on an example circuit with two input factors and one performance parameter

1. The circuit factors are extracted from the input netlist (indicated as *Circuit Factor EXTRACTION* in Figs. 5.14 and 5.15). These factors determine the circuit factor space $\mathcal{F}$.
2. One can sample directly from $\mathcal{F}$, but it is numerically more stable and more accurate to map $\mathcal{F}$ on a normalized and orthogonalized space $\mathcal{F}'$ and to sample from there. Every point $\mathbf{f}_q$ in $\mathcal{F}$ is therefore mapped on a point $\mathbf{f}'_q$ in $\mathcal{F}'$ (indicated as *Circuit Factor NORMALIZATION* in Figs. 5.14 and 5.15):

$$\begin{aligned} \mathbf{f}'_q &= [f'_{q,1}, f'_{q,2}, \ldots, f'_{q,n}, \ldots, f'_N] \\ &\text{with} \quad \mu(f'_{q,n}) = 0 \quad \text{and} \quad \sigma(f'_{q,n}) = 1 \end{aligned} \tag{5.17}$$

   Linear mapping from $\mathcal{F}$ to $\mathcal{F}'$ and vice versa can be found in literature (Montgomery 2008):

$$\mathbf{f}'_q = (\mathbf{f}_q - \mu(\mathcal{F})) \begin{bmatrix} \frac{1}{\sigma(f_1)} & 0 & \ldots & 0 \\ 0 & \frac{1}{\sigma(f_2)} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{1}{\sigma(f_N)} \end{bmatrix} \tag{5.18}$$

   with $\mu(\mathcal{F})$ a vector with the mean values for each factor in $\mathcal{F}$ and $\sigma(f_n)$ the standard deviation on factor f_n. The relationship between $\mathcal{F}'$ and $\mathcal{P}$ is accordingly denoted as ϑ'.
3. A screening DoE defines a set of simulations to identify the most important factors and to detect interactions between different factors (indicated as *SCREENING + 1st-Order MODEL* in Figs. 5.14 and 5.15). The output data of the deterministic reliability simulations result in the development of a first-order model $\hat{\vartheta}_1$. This step has a linear complexity with respect to the number of factors in $\mathcal{F}'$. Each of the experiments, determined by the screening DoE, can be processed in parallel to reduce the simulation time.
4. A residual analysis assesses whether $\hat{\vartheta}_1$, resulting from the screening design, is sufficiently accurate (indicated as *RESIDUAL Analysis* in Fig. 5.14). The model error $\mathbf{e}^i_m$ is given by:

$$\begin{aligned} \mathbf{e}^i_m &= \{e^i_{m,q}\} \\ &= \{P^i_{m,q} - \hat{P}^i_{m,q}\} \quad \text{with} \quad q = \{0, \ldots, Q_{\text{DoE}}\} \end{aligned} \tag{5.19}$$

   where Q_{DoE} represents the number of deterministic reliability simulations. $P^i_{m,q}$ is circuit performance parameter m at time point t_i and for circuit sample q. If the model fits the simulation data well, the errors are small and distributed evenly around zero. A Wilcoxon signed-rank test (Wilcoxon signed rang test 2012) verifies whether the mean error $\mu(e_m)$ for each circuit specification P_m differs significantly from zero, with:

$$\mu(e_m) = \frac{1}{(I+1)(2Q_{\mathrm{DoE}}+3)} \sum_{i=0}^{I} \sum_{q=0}^{2Q_{\mathrm{DoE}}+2} e_{m,q}^{i} \tag{5.20}$$

with I the number of time steps ($t_I = T_{\mathrm{str}}$). Additionally, to test whether the errors are sufficiently small, the standard deviation of the model error $\sigma_{\varepsilon,1}$ is compared to the standard deviation of the simulation data itself σ_{P}. If $\sigma_{\varepsilon,1} < 0.1\sigma_{\mathrm{P}}$, the model error is considered sufficiently small. If one or both tests fail, extra simulations are needed in order to get a more complex second-order model $\hat{\vartheta}_2$. This approach now ensures a good model accuracy, but still guarantees a limited simulation complexity.

5. $\hat{\vartheta}_1$ models the behavior of a (deterministic) computer code. Lack of fit, possibly detected in the previous step, is therefore entirely related to modeling errors (i.e. an incomplete set of regression terms) and not to measurement errors or noise. To reduce these errors, extra simulations and regression terms are needed. A regression design is used to finetune $\hat{\vartheta}_1$ with some additional terms, resulting in $\hat{\vartheta}_2$ (indicated as *REGRESSION + 2nd-Order MODEL* in Figs. 5.14 and 5.15). The latter step has an exponential complexity in terms of the number of factors. However, the goal of this work is not to obtain a highly accurate model, but to analyze the impact of process variations and transistor aging, on the performance of a circuit, in a reasonably short time. The regression DoE is therefore only executed when needed.
6. If circuit specifications are provided, the TTF distribution is calculated (indicated as *TTF Distribution GENERATION* in Figs. 5.14 and 5.15). This is done via a Monte-Carlo analysis on $\hat{\vartheta}_2$. The evaluation of that model is very fast, compared to an actual circuit reliability analysis. Therefore a very large number of Monte-Carlo samples can be evaluated. For every time point t_i, the simulator evaluates whether each sample meets the circuit specifications. Samples that do not meet all the specifications are labeled as failures. Eventually, the number of failures at each time point determines the TTF distribution. Given the large number of Monte-Carlo simulations, the confidence interval on the obtained result mainly depends on the model error. This error is calculated as defined in Eq. (5.19).

At the output of the simulator, the user receives $\hat{\vartheta}_2$, which models the circuit performance as a function of the most important input factors and as a function of the stress time. If circuit specifications are given, an estimation for the circuit time-to-failure distribution is also provided. More details on the different parts of the simulator are explained in the next sections.

Screening Design

The screening design is the first DoE in the stochastic reliability simulation flow. The objectives of this DoE (also see Fig. 5.14) are:

1. To analyze the individual impact of all circuit factors (i.e. the dimensions of $\mathcal{F}$) on the circuit performance space $\mathcal{P}$.
2. To find a first-order model $\hat{\vartheta}_1$ for ϑ (also see Eq. (5.13)).

The screening design used in this work is a systematic fractional replicate design (SFRD) and assumes a system that can be modeled by a linear function (Cotter 1979). The system ϑ, that is modeled here, is a function of the input factors $\mathbf{f}_q$ and the stress time t_i. The output is a vector of circuit performance parameter values $\mathbf{P}_q^i$:

$$\begin{aligned}\vartheta(\mathbf{f}_q, t_i) &= \mathbf{P}_q^i \\ &= [\vartheta_{1,q}^i, \ldots, \vartheta_{m,q}^i, \ldots, \vartheta_{M,q}^i] \\ &= [P_{1,q}^i, \ldots, P_{m,q}^i, \ldots, P_{M,q}^i]\end{aligned} \tag{5.21}$$

with M the number of circuit performance parameters. Further, $q = \{1, \ldots, Q\}$ with Q the number of experiments in the screening design. At each time point t_i, for each circuit performance parameter P_m, a separate model need to be build. When written as a linear function of the normalized input factors these models $\hat{\vartheta}_m'^i$ are:

$$\begin{aligned}\hat{\vartheta}_m'^i &= \hat{P}_m^i \\ &= a_0 + \sum_{n=1}^{N} a_n f_n' + \sum_{n=1}^{k-1} \sum_{k=2}^{N} a_{nk} f_n' f_k' + \cdots + a_{1\ldots N} f_1' \ldots f_N'\end{aligned} \tag{5.22}$$

where $\{a_n, a_{nk}, \ldots\} = f(m, i)$ are the model coefficients and N is the number of circuit factors (i.e. the number of dimensions in $\mathcal{F}$).[5, 6]

Every screening design DoEs consists of $2N+3$ experiments (see Table 5.2), each of which corresponds to a point in $\mathcal{F}'$. Each experiment is evaluated by the deterministic reliability simulator described in Sect. 5.2, eventually resulting in a set of outputs for each time point t_i : $\mathbf{P}_q^i$, $q = \{1, \ldots, 2N+3\}$. To model $\hat{P}_m^i$ each parameter $a_n, n = \{1, \ldots, N\}$, can be calculated as:

$$a_n = \frac{1}{4}\left[(P_{m,2N+1}^i - P_{m,N+n}^i) + (P_{m,n}^i - P_{m,0}^i)\right] \tag{5.23}$$

where $P_{m,n}^i$ represents the simulation result for the nth screening experiment. The sum of the first-order interaction effects of each factor f_n' with any other factor f_k', $k \neq n$, can also be estimated:

---

[5] The model parameters are different for each time point and for each performance parameter. This is not indicated explicitly here to improve readability.

[6] $\hat{\vartheta}_m'^i$ uses the normalized factor space $\mathcal{F}'$ as input (see Eq. (5.17)) and the DoE is also designed in $\mathcal{F}'$. To evaluate the response of a point in $\mathcal{F}$, an extra normalization step, using Eq. (5.18), is required.

Table 5.2 Systematic fractional replicate screening design setup

$\text{DoE}_S \in \mathbb{R}^{(2N+3)\times N} = [\mathbf{f}'_1; \ldots; \mathbf{f}'_{2N+3}]$
$\text{DoE}_S[0, :] = [-1, \ldots, -1]$
$\text{DoE}_S[1 : N, :] = \begin{bmatrix} 1 & -1 & \ldots & -1 \\ -1 & 1 & \ldots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \ldots & 1 \end{bmatrix}$
$\text{DoE}_S[N+1 : 2N, :] = \begin{bmatrix} -1 & 1 & \ldots & 1 \\ 1 & -1 & \ldots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \ldots & -1 \end{bmatrix}$
$\text{DoE}_S[2N+1, :] = [1, \ldots, 1]$
$\text{DoE}_S[2N+2, :] = [0, \ldots, 0]$

$$\sum_{k\neq n} a_{nk} = \frac{1}{4}\left[(P^i_{m,2N+1} - P^i_{m,N+n}) - (P^i_{m,n} - P^i_{m,0})\right] \tag{5.24}$$

If, for a given factor f'_n, both a_n and the sum of interactions with any other factor $\sum_{k\neq n} a_{nk}$ are small, this factor will most probably not have a significant impact on the circuit behavior and can therefore be neglected. Also, since Eq. (5.23) gives an estimation for the main impact of each factor on the circuit output, a first-order model $\hat{\vartheta}'^i_{1,m}$ for ϑ'^i_m can be found:

$$\hat{\vartheta}'^i_{1,m} = \hat{P}^i_m = a_0 + \sum_{n=1}^{N} a_n f'_n \tag{5.25}$$

$$\text{with } a_0 = \frac{1}{2N+3}\sum_{n=0}^{2N+2} P^i_{m,n}$$

The two objectives stated at the beginning of this section are therefore fulfilled: the screening design enables identification of the most important factors and a first-order model $\hat{\vartheta}'^i_{1,m}$ can be derived. Nevertheless:

1. Any factor with a large interaction with another factor and an equally large, but opposite, interaction with a third factor, will wrongfully be neglected.
2. $\hat{\vartheta}'^i_{1,m}$ is only a first-order model and does not include nonlinear effects or interaction effects between different factors.

The first problem is very unlikely to occur in real circuits. Further, to verify this issue a full factorial design is required (Montgomery 2008). Such an experimental design is computationally very expensive and is therefore not considered here. To solve

the second problem, a center design was added to the screening design (i.e. $\text{DoE}_S[2n+2,:] = \mathbf{f}'_{2N+3} = [0,\ldots,0]$ in Table 5.2), enabling detection of nonlinear input-output behavior. A residual analysis (see Fig. 5.14) assesses whether the lack of fit between $\hat{\vartheta}'^{i}_{1,m}$ and ϑ'^{i}_{m} is sufficiently small. If needed extra simulations are executed, defined by a subsequent regression design and explained in the next section.

Regression Design

If the screening design, discussed in the previous section, does not yield a sufficiently accurate model for ϑ, a regression design is needed. The extra simulations, defined by this regression design, are used to extract information on nonlinearities and interaction effects. In contradiction to the screening design, the number of experiments required to do this grows more than linearly with the number of factors. Therefore, to minimize the computation time, only factors with a significant impact on the circuit behavior are studied. These are selected based on the results of the screening design (see Eqs. (5.23) and (5.24)):

$$f'_{d,d\in\{1,\ldots,N\}} = \text{dominant} \quad \Updownarrow \quad \exists\, \hat{\vartheta}'^{i}_{1,m} : \left| a_d + \sum_{k\neq d} a_{dk} \right| \geq \alpha \max_{n\in\{1,\ldots,N\}} \left(\left| a_n + \sum_{k\neq n} a_{nk} \right| \right) \tag{5.26}$$

with parameter $0 < \alpha < 1$ and typically $\alpha = 0.1$. The number of dominant factors f'_d satisfying Eq. (5.26) is D. Typically, for circuits with $N > 100$, $D < N/10$. Factors that do not satisfy Eq. (5.26) are set to their mean value 0, for all regression simulations here. Factors f'_d are set according to the regression design.

The regression design used here is a central composite design. Such a design is often used to develop a second-order (quadratic) model of a system and requires far less experiments when compared to a full three-level factorial experiment that is otherwise required (Montgomery 2008). For example, when studying 10 factors, the former requires 149 experiments, while a three-level factorial design consists of 59049 experiments. The DoE consists of three distinct sets of experiments:

1. A *factorial and fractional factorial design*. These designs are regularly used in industrial experimentation and regression modeling. Two-level factorial designs are very useful for testing or estimating linear and interaction effects, while fractional factorial designs do the same with fewer experiments (Montgomery 2008; Engineering statistics handbook 2012). Here, a resolution V fractional factorial (R5FF) design is used. This DoE allows to identify both the linear effect for every factor (indicated as a_n in Eq. (5.22)), as well as all first-order interaction effects (indicated as a_{nk} in Eq. (5.22)). According to (Sanchez and Sanchez 2005) an R5FF can be generated for any number of factors when using a Hadamard-ordered

Table 5.3 Column indices in a Hadamard matrix to create a resolution V fractional factorial design

# Factors	# Design points	Column indices		
1	2	1		
2	4	2		
3	8	4		
4–5	16	8	15	
6	32	16		
7–8	64	32	51	
9–11	128	64	85	106

The index of the first column is 0. For an experiment with D factors, the first D indices in the table define the columns for the DoE

matrix. This matrix H_v is defined as:

$$H_0 = (1) \quad and \quad H_{v+1} = \begin{pmatrix} H_v & H_v \\ H_v & -H_v \end{pmatrix} \tag{5.27}$$

A R5FF design for the D most dominant factors (see Eq. (5.26)) consists of a subset of the columns of H_v. Each column in that subset defines the values for one factor in the R5FF design, the number of experiments in the DoE are given by the number of elements in one column. Table 5.3 lists the column indices for a R5FF design for up to 11 factors (the index of the first column is 0). For example, for a circuit with 4 factors, the columns with indices {1, 2, 4, 8} of a H_4 Hadamard matrix represent a R5FF design. The first row of each column defines the factor values for the first experiment, the second row defines the second experiment, etc.

2. A *center design*. Here, all the factors in the design are set to their nominal value: 0. However, since this design is already part of the screening design, this is not included here.
3. A set of *axial points*. To capture weakly nonlinear effects, the regression design is augmented with DoE_A, which requires $2D$ extra simulations. Each of the experiments in DoE_A are identical to the center design except for one factor, which takes in turn a value below and above the nominal value. The setup of these experiments and the corresponding outputs is listed in Table 5.4.

The simulation results of each experiment defined by central composite design are combined with the results of the screening design and used to create a set of second-order models $\hat{\vartheta}'^{i}_{2,m}$ for each performance parameter and at each time point. This model will be discussed in the next section.

Stochastic Circuit Model

The simulation results obtained from both experimental designs, as discussed above are used to build a model $\hat{\vartheta}'$ for ϑ'. At each time point t_i and for each circuit

Table 5.4 Axial design setup

$$\mathrm{DoE_A} \in \mathbb{R}^{(2D)\times D}$$

$$\mathrm{DoE_A}[0 : D-1, :] = \begin{bmatrix} D^{1/4} & 0 & \dots & 0 \\ 0 & D^{1/4} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & D^{1/4} \end{bmatrix}$$

$$\mathrm{DoE_A}[D : 2D, :] = \begin{bmatrix} -D^{1/4} & 0 & \dots & 0 \\ 0 & -D^{1/4} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & -D^{1/4} \end{bmatrix}$$

performance parameter P_m a separate 2nd-order model is built:

$$\hat{\vartheta}'^{i}_{2,m} = \hat{P}^i_m \approx \vartheta'^{i}_m = P^i_m \tag{5.28}$$

In this work different regression methods have been explored to build the model:

1. *Physical model*: based upon approximate modeling of physical phenomena within the circuit.
2. *Empirical model*: an analytical model to adequately fit a data set:
   (a.) Ordinary least squares (OLS): a widely applied technique, limiting the sum of squared residuals for a pre-defined polynomial model.
   (b.) Radial basis function (RBF) neural network: a system with one hidden layer of artificial neurons, using the input data set as training data.
3. *Tabular model*: a model evaluated through interpolation or extrapolation of a data set in a look-up table. Locally weighted scatterplot smoothing (LOESS), for example, fits simple OLS models to localized subsets of the input data set.

A physical model can be very accurate, even if only a small input data set is available, but it requires a thorough knowledge of the circuit and is very circuit-dependent. In this work, the focus is on developing a reliability simulator requiring as little input and knowledge from the user as possible. Therefore, a physical model is not a good option here. A tabular model can be very accurate but requires user-defined and circuit dependent smoothing parameters. Combined with a fairly large model evaluation time and poor extrapolation capabilities, tabular models are therefore also not a good option here. After implementation and evaluation, an RBF method showed a better behavior when modeling highly nonlinear functions, but performs worse when modeling functions with a large correlation coefficient R^2 (i.e. linear or weakly nonlinear functions). The accuracy of an RBF also deteriorates fast when an increasing number of dimensions and an increasing number of data values are fitted (Hussain et al. 2002; Zhang et al. 2005). Here, the number of dimensions is typically large (larger than 100 for a typical analog circuit). Further, the deviation from the

nominal design point is rather small (i.e. a function of relatively small statistical variations of process and aging-related parameters). A polynomial model fitted with OLS and consisting of an average term, linear terms for each factor and quadratic and interaction terms to model weak nonlinearities, therefore typically results in a good model for the intended purpose:

$$\hat{\vartheta}'^{i}_{2,m} = a_0 + \sum_{n=1}^{N} a_n f'_n + \sum_{n=1}^{k-1} \sum_{k=2}^{N} a_{nk} f'_n f'_k + \sum_{n=1}^{N} a_{n,2} f'^{2}_n \tag{5.29}$$

The simulation complexity and accuracy of the proposed technique are discussed in more detail in the next section.

Complexity and Accuracy

Above, the DoE-based stochastic reliability simulator has been explained in detail. This section studies the simulation complexity, accuracy and speed of the proposed method. Also, the results are compared to the MC-based approach discussed in Sect. 5.3.2.

The simulation accuracy mainly depends on:

1. The accuracy of the transistor aging compact models.
2. The accuracy of the deterministic reliability simulator.
3. The accuracy of the stochastic circuit model $\hat{\vartheta}$.

The simulation accuracy of the deterministic reliability simulator mainly depends on the correctness of the transistor aging compact models. The accuracy of the deterministic reliability simulator itself has been discussed in Sect. 5.2.2. Further, a set of compact models, suited for analog circuit reliability simulation, have been proposed in Chap. 3. These models have been validated with measurements. Each commercial or academic reliability simulator, reviewed in Chap. 4, and the MC-based stochastic simulator discussed in Sect. 5.3.2, have to deal with this problem and require accurate device compact models to yield good results. In this section, the focus is on the accuracy of the $\hat{\vartheta}$, the circuit model that is generated based on a set of specific experimental designs (see Eq. (5.29)).

The DoE-based method is tested on a set of seven common analog and digital circuits. The circuits and the observed performance parameters are listed in Table 5.5. Also, the number of factors for each test circuit is depicted. This number equals the number of dimensions in $\mathcal{F}$ and corresponds to the design complexity. To build the circuit model, each factor f_n was varied in a three sigma range:

$$\mu(f_n) - 3\sigma(f_n) \leqslant f_n \leqslant \mu(f_n) + 3\sigma(f_n) \tag{5.30}$$

Table 5.5 also shows the number of deterministic reliability simulations, needed to build the circuit model (indicated as #DRS). These training samples are defined by

Table 5.5 DoE-based stochastic reliability simulation: accuracy validation

	#Factors	#DRS	$\mu(\varepsilon_{\text{rel}})$ (%)	$\sigma(\varepsilon_{\text{rel}})$ (%)	$\hat{\sigma}(\varepsilon_{\text{rel}})$ (%)	$\frac{\sigma(\text{P})}{\sigma(\varepsilon_{\text{abs}})}$	Time (min:s)
1	5	19	0.04	0.57	0.65	45.9	1:28
2	14	43	−0.31	1.44	1.19	16.2	2:32
3	15	53	−0.09	0.28	0.46	74.0	3:37
4	27	95	−0.20	0.41	0.40	47.4	4:46
5	35	121	0.14	1.30	1.01	12.6	9:09
6	20	61	0.01	0.26	0.26	94.8	2:06
7	141	343	−0.12	0.98	1.10	22.3	9:46

The experiments have been executed on a dual-quad core 2.8 GHz Intel Xeon processor with 8 GB of RAM
1: One-stage amplifier (gain)
2: LC-VCO (oscillation amplitude)
3: Differential pair amplifier (gain)
4: Symmetrical OTA (offset)
5: Ring oscillator (oscillation frequency)
6: AND gate (fall time)
7: IDAC (output voltage)

the screening and regression DoE as explained above. Each circuit was simulated over a stress time T_{str} of four months. The resulting circuit model $\hat{\vartheta}$ (for $t \leq T_{\text{str}}$) was then evaluated in 500 test samples $\mathbf{f}_q$, $q = \{1, \ldots, 500\}$, uniformly distributed over the factor space $\mathcal{F}$. Each test sample was also evaluated using the deterministic reliability simulator yielding a 'gold standard' value $\vartheta(\mathbf{f}_q)$, which can be compared with the output given by $\hat{\vartheta}(\mathbf{f}_q)$[7]:

$$\begin{aligned} \varepsilon_{\text{abs},q} &= \hat{\vartheta}(\mathbf{f}_q) - \vartheta(\mathbf{f}_q) \\ &= \hat{P}_q - P_q \end{aligned} \tag{5.31}$$

To compare model errors for the different circuits, each model error was calculated relative to the respective circuit parameter range:

$$\varepsilon_{\text{rel},q} = \frac{\hat{P}_q - P_q}{\max\limits_q (P_q) - \min\limits_q (P_q)} \tag{5.32}$$

For each test circuit, the mean model error $\mu(\varepsilon_{\text{rel}})$ and the error standard deviation $\sigma(\varepsilon_{\text{rel}})$ are listed in Table 5.5. The table also depicts $\sigma(\text{P})/\sigma(\varepsilon_{\text{abs}})$, with $\sigma(\text{P})$ the circuit performance parameter standard deviation and $\sigma(\varepsilon_{\text{abs}})$ the absolute model error standard deviation. When this ratio is larger than 10, which is true for all test

[7] The error can be calculated for each circuit performance parameter P_m, at each time point t_i. To improve readability, this is not indicated explicitly in the equations. To obtain the model errors depicted in Table 5.5, the error at time T_{str} was taken.

circuits, the circuit model is considered to be sufficiently accurate. This corresponds to the criterion used above to determine whether the first-order model returned from the screening analysis is sufficiently accurate. In addition, Fig. 5.16 depicts the model errors for each circuit on a normal probability plot. For each circuit the model errors are evenly distributed around 0, indicating a good model fit over the studied factor range. Finally, Table 5.5 lists $\hat{\sigma}(\varepsilon_{\text{rel}})$, an estimate for the relative model error standard deviation, based on a cross validation of the model (Davison and Hinkley 1997). Cross validation is a technique to assess the accuracy of a predictive model. The model is built for a subset of the original input data set (called the training set) and validated on the remaining datasamples (the testing or validation set). This routine is done multiple times. The error between the cross-validation models and the testing sets is a measure for the expected error in a practical situation.[8] Indeed, a designer will, besides the evaluation of the model, also be interested in the error on each evaluation. In Fig. 5.16, the estimated model error $\hat{\varepsilon}_{\text{rel}}$ is indicated with a solid line. For all circuits under test, $\hat{\varepsilon}_{\text{rel}}$ coincides fairly well with the actual error distribution ε_{rel} and is therefore a good estimate for the latter.

To build the circuit model, the required number of deterministic reliability simulations increases linearly with the number of circuit factors (O(N)) for the screening DoE, while for the regression DoE it increases exponentially (O(2^N)). The main reason to include DoE techniques in the stochastic reliability simulator, however, is to limit the number of deterministic reliability simulations and to minimize the simulation time. Therefore, the regression design is only used when necessary (also see Eq. (5.26)). Eventually, this results in a nearly-linear overall simulation complexity of the proposed approach. This is clear from Table 5.5, where #factors/#DRS is approximately constant. In addition, the simulation complexity is also plotted in Fig. 5.17. This figure shows the required number of deterministic reliability simulations, as a function of the design complexity, for all test circuits. On the figure, it is also clear how the computational effort increases linearly with the number of factors in the circuit. Nevertheless, the circuit model not only includes the main linear effects, but also encloses the most important nonlinear and interaction effects.

Finally, to validate the simulation speed of the proposed method, the circuit yield at time $t = 0$s was calculated with both the proposed simulator as well as with the MC-based simulator discussed in Sect. 5.3.2. The simulation time for both yield calculations can only be compared if the accuracy on the yield prediction is the same. The variance on the yield, when calculated with a MC-based stochastic reliability simulator is defined as:

[8] Cross validation is similar to the bootstrapping method, used to estimate the error on the results for the MC-based reliability simulator in Sect. 5.3.2. Both methods use resampling to estimate sample statistics (in this case the variation) and yield similar results. Bootstrapping, however, first estimates the entire distribution and then calculates the sample statistic, while cross validation only provides estimates for the variance on the point estimator. Bootstrapping is therefore more computer intensive and not used here to minimize simulation times.

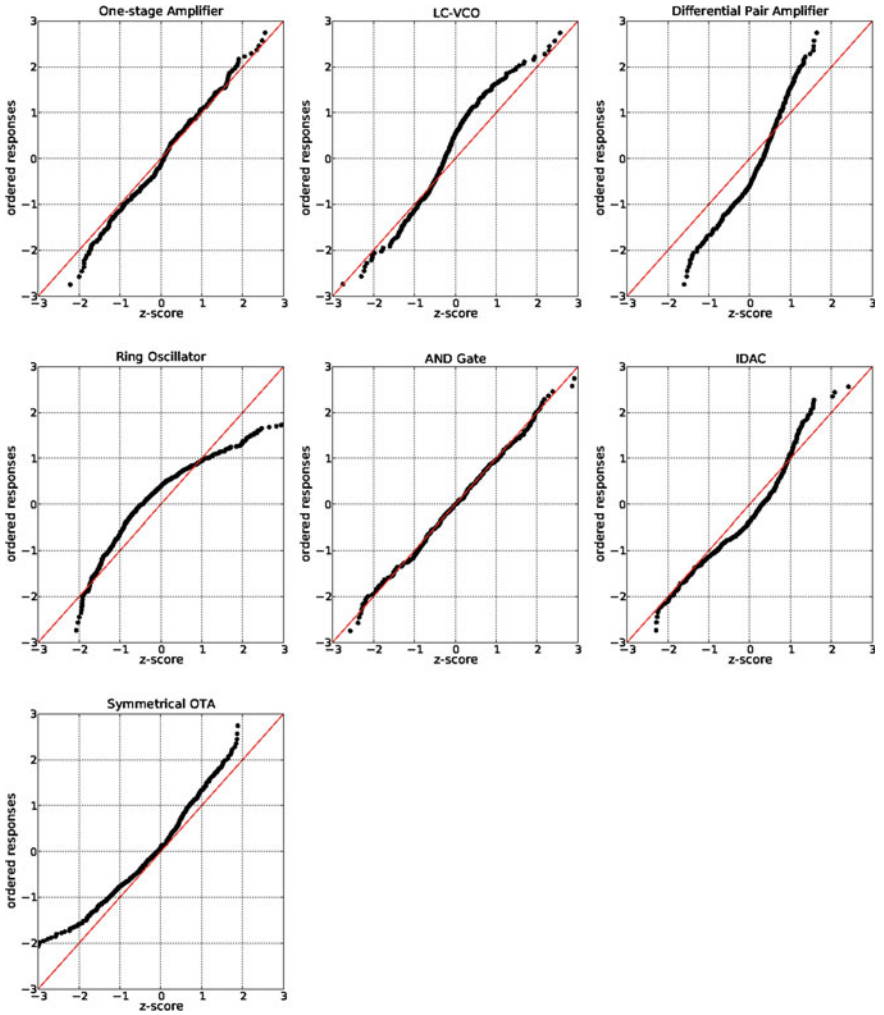


Fig. 5.16 Normal probability plot of the actual (*dotted line*) and estimated (*solid line*) model error, for all test circuits

$$\sigma^2(Y_{\mathrm{MC}}) = \frac{Y_{\mathrm{MC}}(1 - Y_{\mathrm{MC}})}{Q_{\mathrm{MC}}} \tag{5.33}$$

with Q_{MC} the number of simulations and Y_{MC} the yield value (Elias 1994). When using a DoE-based stochastic reliability simulation, the simulator first builds a model $\hat{\vartheta}$ that includes the most important process variability and stochastic aging effects. Then, the yield value Y_{DoE} is calculated with a Monte-Carlo simulation on $\hat{\vartheta}$. The total variance on Y_{DoE} is:

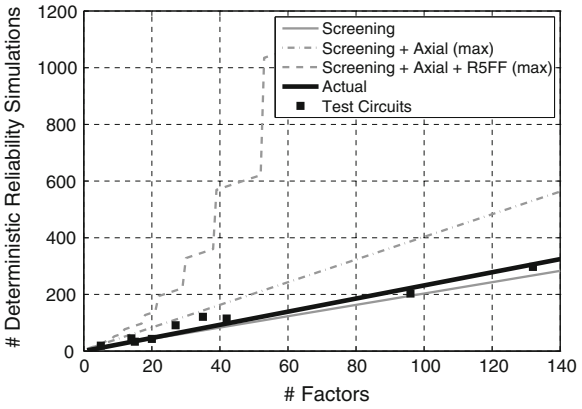


Fig. 5.17 The simulation complexity (required number of deterministic reliability simulations to build the circuit model) as a function of the number of circuit factors. The proposed method has a nearly-linear simulation complexity

Table 5.6 Simulation speedup for yield calculations with the DoE-based method

	$\hat{Y}_{\mathrm{DoE}}$ (t = 0) (%)	$\sigma^2(Y_{\mathrm{DoE}})$ (t = 0) (%)	Speedup
1	97.6	0.12	461
2	99.7	0.09	44
3	99.5	0.02	1260
4	99.8	0.03	697
5	99.7	0.05	30
6	99.3	0.04	536
7	99.8	0.05	22

1: One-stage amplifier (gain)
2: LC-VCO (oscillation amplitude)
3: Differential pair amplifier (gain)
4: Symmetrical OTA (offset)
5: Ring oscillator (oscillation frequency)
6: AND gate (fall time)
7: IDAC (output voltage)

$$\sigma^2(Y_{\mathrm{DoE}}) = \sigma^2(Y_{\mathrm{MC}}) + \sigma^2(Y_{\varepsilon}) \tag{5.34}$$

with $\sigma^2(Y_{\varepsilon})$ the variance on the yield due to the model error. The latter is only a function of the model accuracy and will therefore dominate $\sigma^2(Y_{\mathrm{DoE}})$ if Q is large. To estimate $\sigma^2(Y_{\mathrm{DoE}})$, for each test circuit, a Monte-Carlo simulation with 1e5 samples (Q = 1e5), using $\hat{\vartheta}$, was conducted. Any remaining variance on the circuit yield was assumed to originate exclusively from the model error.

For all test circuits, the yield and the estimated standard deviation $\sigma^2(Y_{\mathrm{DoE}})$ are listed in Table 5.6. The number of simulations needed to obtain the same accuracy with a standard Monte-Carlo reliability simulation can be derived from Eq. (5.33):

$$Q_{\mathrm{MC}} = \frac{Y_{\mathrm{DoE}}(1 - Y_{\mathrm{DoE}})}{\sigma^2(Y_{\mathrm{DoE}})} \tag{5.35}$$

The yield simulation speedup is then defined as:

$$\mathrm{Speedup} = \frac{Q_{\mathrm{MC}}}{Q_{\mathrm{DoE}}} \tag{5.36}$$

with Q_{DoE} being the number of deterministic reliability simulations needed to generate the RSM model (also see Table 5.5). The speedup is listed in Table 5.6 and varies between 22 and 1260×, i.e. between 1 to 3 orders of magnitude, for the circuits under test.

5.3.4 Circuit Example

In Sect. 5.2, the deterministic reliability simulator has been demonstrated on an example comparator circuit. Here, the same example circuit is used to demonstrate the stochastic reliability simulator discussed in this section. Adding the impact of stochastic effects reveals a lot of extra information about the time-dependent performance of the comparator. In the sections below, the circuit schematic is briefly reviewed. Then, the impact of process variations and stochastic aging effects on the circuit offset is calculated by hand using the first-order model proposed in Sect. 3.6. Finally, the simulation results from the stochastic reliability simulator are presented and discussed.

Circuit Schematic

The example circuit is the same as the demonstrator circuit for the deterministic reliability simulator. For convenience, the circuit schematic is repeated in Fig. 5.18. The comparator is subjected to the same stress voltages as before. At the input of the comparator, a sine wave with an amplitude of 0.4 V and a DC bias of 0.5 V was applied. The voltage at the reference input is again 0.2 V (also see Fig. 5.6). The circuit performance parameters also remain the same: the input offset voltage and the slew rate. The average impact of transistor aging on the circuit behavior is therefore expected to be similar to the results obtained for the deterministic simulation (see Sect. 5.2.3). Adding stochastic effects, however, will reveal extra information about the time-dependent spread on that average. Model parameters for the process variations and stochastic aging effects are obtained from (Lewyn et al. 2009) and (Kaczer et al. 2010) respectively.

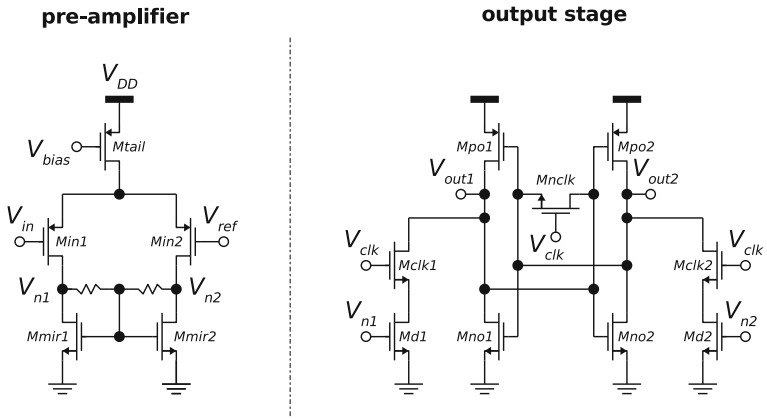


Fig. 5.18 The clocked comparator, used as a demonstrator circuit for the stochastic reliability simulator

Hand Calculations

In Sect. 5.2.3, the average aging-induced offset shift was calculated with the first-order model proposed in Sect. 3.6. Here, the same model is used to calculate the spread on the offset due to process variations and stochastic aging effects. The variance on the offset is derived from Eq. (5.6) and can be written as:

$$\sigma^2(V_{\text{offset}}) = \sigma^2(\delta V_{\text{TH,Min}}) + \frac{\sigma^2(\delta V_{\text{TH,Md}})}{A^2_{\text{pre}}} \tag{5.37}$$

$$\approx \sigma^2(\delta V_{\text{TH0,Min}}) + \sigma^2(\Delta V_{\text{TH,Min2}}) + \frac{\sigma^2(\delta V_{\text{TH0,Md}})}{A^2_{\text{pre}}} + \frac{\sigma^2(\Delta V_{\text{TH,Md2}})}{A^2_{\text{pre}}} \tag{5.38}$$

with $\sigma^2(\delta V_{\text{TH0,Min}})$ and $\sigma^2(\delta V_{\text{TH0,Md}})$ the variance on the initial mismatch between Min1 and Min2, and Md1 and Md2, respectively. $\sigma^2(\Delta V_{\text{TH,Min2}})$ and $\sigma^2(\Delta V_{\text{TH,Md2}})$ are the variance on the shift in V_{TH} for Min2 and Md2 due to stochastic aging effects. The aging-induced variance on the V_{TH} of Min1 and Md1 is not included in (5.39) because these transistors are, according to the hand calculations, not degrading (also see Sect. 5.2.3). Further, the aging-induced variance on the V_{TH} of a transistor is a function of the average V_{TH} shift (also see Sect. 3.6):

$$\sigma^2(V_{\text{TH}}) = \underbrace{\frac{A^2_{\text{VTH}}}{WL}}_{\text{process variations}} + \underbrace{\frac{A^2_{\text{BTI}}|\Delta V_{\text{TH}}|}{WL}}_{\text{stochastic aging effects}} \tag{5.39}$$

with A_{VTH} and A_{BTI} process-dependent parameters. Typical values for a 32 nm process are $A_{\text{VTH}} = 1.7\text{e-}9$ and $A_{\text{BTI}} = 5.7\text{e-}9$. Equation (5.39) can be recast to:

$$\frac{\sigma(\Delta V_{\mathrm{TH}})}{\sigma(V_{\mathrm{TH0}})} = \frac{A_{\mathrm{BTI}}\sqrt{|\Delta V_{TH}|}}{A_{\mathrm{VTH}}} \approx 3.35\sqrt{|\Delta V_{\mathrm{TH}}|} \quad \text{(for a 32 nm CMOS process)} \tag{5.40}$$

The calculated aging-induced V_{TH} shift of Min2 and Md2 is very small ($\Delta V_{\mathrm{TH,Min2}} = 1.9\,\mathrm{mV}$ and $\Delta V_{\mathrm{TH,Md2}} = 10.7\,\mathrm{mV}$, also see Table 5.1). Therefore, using equations (5.40) and (5.39), the anticipated increase in $\sigma^2(V_{\mathrm{offset}})$ due to aging is also very small and will primarily depend on the initial transistor mismatch:

$$\sigma^2(V_{\mathrm{offset}}) \approx \sigma^2(\delta V_{\mathrm{TH0,Min}}) + \frac{\sigma^2(\delta V_{\mathrm{TH0,Md}})}{A^2_{\mathrm{pre}}} \tag{5.41}$$

As calculated in Sect. 5.2.3, the expected shift in slew rate is negligible. Therefore, the variance on the slew rate will also primarily depend on the initial process-related variations:

$$\sigma^2(\mathrm{SR}) = \mathrm{SR}^2\left(\frac{4\sigma^2(V_{\mathrm{TH}})}{V^2_{\mathrm{GST}}} + \frac{4\sigma^2(C_{\mathrm{L}})}{C^2_{\mathrm{L}}}\right) \tag{5.42}$$

with $\sigma^2(C_{\mathrm{L}})$ the variance on the load capacitor C_{L}. In conclusion, from the hand calculations the variability on the circuit performance parameters is assumed to remain pretty constant over time.[9] The average value of the circuit performance parameters, however, will change over time. In the next section, these assumptions are verified with the stochastic reliability simulator.

Simulation Results

To simulate the impact of stochastic aging effects on the performance of the example circuit, one can either use the MC-based simulator (see Sect. 5.3.2) or the DoE-based simulator (see Sect. 5.3.3). As shown in Sect. 5.3.3, both simulators yield similar results. The DoE-based simulator, however, is one to three orders of magnitude faster (also see Table 5.6). Therefore, to obtain the results below, the DoE-based simulator was primarily used. However, a limited MC-based stochastic simulation was also conducted and the results were compared to the ones obtained with the DoE-based simulator.

Figure 5.19 shows the comparator input offset as a function of time, evaluated on 100 random samples evaluated with a circuit model generated with the DoE-based reliability simulator. As expected from the hand calculations, the mean offset shift corresponds to the result obtained from the deterministic reliability simulator (shown in Fig. 5.7). In addition, Fig. 5.19 can be used to get an first impression of the initial

[9] Other circuit examples where the performance parameter variability does change over time will be discussed in Chap. 6.

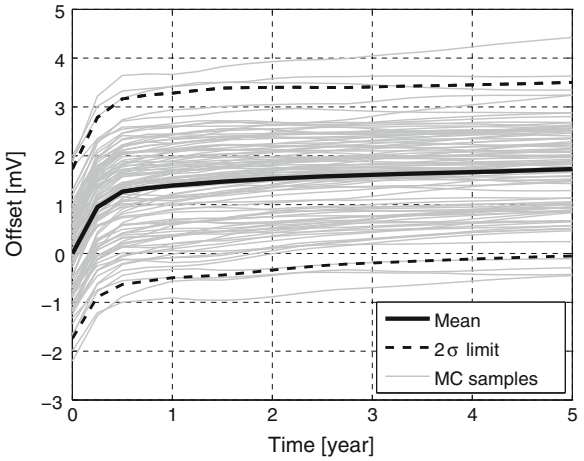


Fig. 5.19 Comparator input offset as a function of the stress time for 100 random samples evaluated with a circuit model generated with the DoE-based reliability simulation method

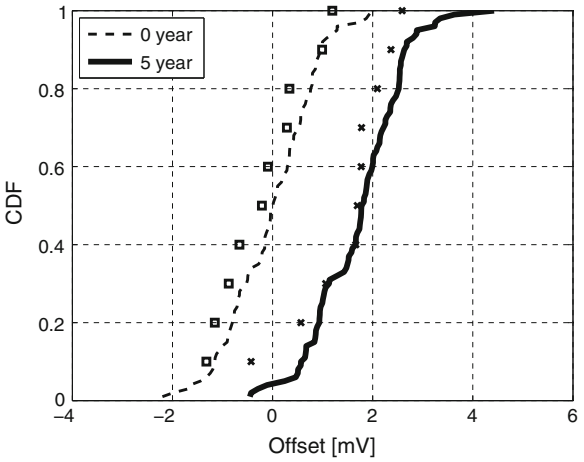


Fig. 5.20 Cumulative distribution function of the offset for a fresh circuit (time t = 0 s) and for an aged circuit (time t = 5 years). A circuit model, generated by the DoE-based stochastic simulator, was evaluated in 100 random sample points (*lines*), while 10 other random samples were evaluated with the MC-based stochastic simulator (*markers*)

spread on the offset due to process variations, as well as the time-dependent shift due to deterministic and stochastic aging effects.

Next, Fig. 5.20 depicts a cumulative density function (CDF) of the offset for a fresh circuit (time t = 0 s) and an aged circuit (time t = 5 year). In essence, this plot shows the same information as Fig. 5.19 but it is easier to extract numerical values. The

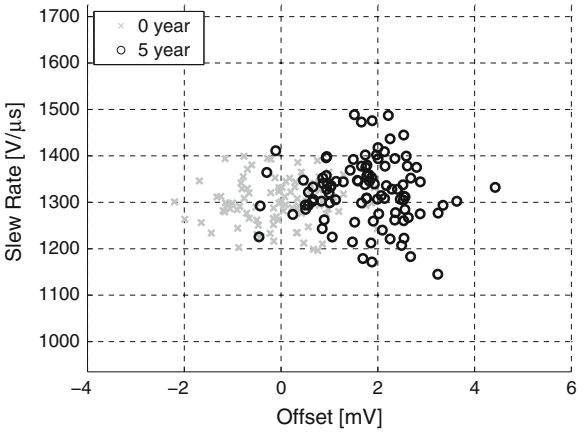


Fig. 5.21 The circuit performance space for the circuit offset and the slew rate for a fresh circuit and after a stress time of five years for the example comparator circuit in a 32 nm CMOS technology

solid lines in Fig. 5.20 represent the evaluation of the circuit model, generated by the DoE-based stochastic simulator, in 100 sample points. The markers are 10 (different) sample points evaluated with the MC-based stochastic simulator. The results match very well, demonstrating a good accuracy of the circuit model. Although the average offset value shifts due to transistor aging, the spread on that value (which is a function of the slope of the CDF) does not change much. This observation again corresponds to the result obtained from hand calculations in the previous section.

The circuit performance space $\mathcal{P}$, for the two observed performance parameters for a fresh circuit and after a stress time of five years, is depicted in Fig. 5.21. The average slew rate remains more or less constant, which is in agreement with the deterministic simulator results depicted in Fig. 5.7. The offset, however, shifts a lot. The average offset shift increases to nearly 2 mV, while some outlier samples even reach an offset value larger than 4 mV. From this figure, it is clear that if a designer does not want this circuit to fail after 5 years, he or she should either include redundancy or design the circuit taking into account the combined variation visualized by the two sample clouds in Fig. 5.21.

Finally, Fig. 5.22 shows the predicted time-to-failure distribution with as failure criterion $|V_{\text{offset}}| > 2.5\,\text{mV}$. A 95 % confidence bound on that prediction is also given. Due to the logarithmic BTI time-dependence, 10 % of the samples fails before one year, while after 5 years of stress another 12 % has failed. The average offset shift, also obtained from the deterministic reliability simulation, does not surpass the 2.5 mV limit. Nevertheless, due to process variations and stochastic aging effects, part of the samples do fail before the intended circuit lifetime. This result demonstrates the importance of using a stochastic reliability simulation instead of a deterministic simulation only.

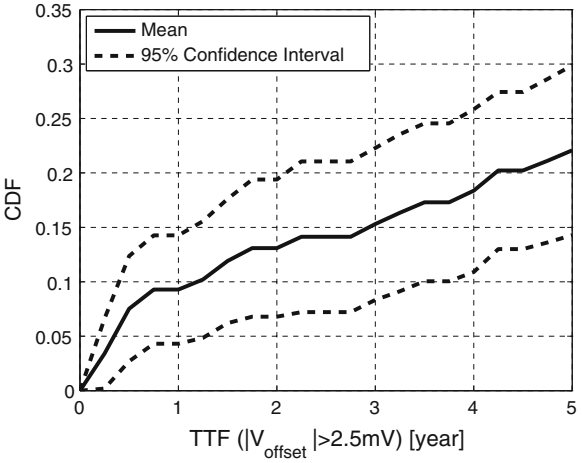


Fig. 5.22 The cumulative density function for the time to failure. A circuit is considered to fail if the input offset surpasses 2.5 mV. A 95 % confidence interval is also given

5.4 Hierarchical Reliability Simulation

Section 5.3 has discussed the development of a stochastic reliability simulator. Such a simulator analyzes the impact of deterministic and stochastic unreliability effects on the time-dependent performance of a circuit. An efficient implementation of this simulator, using design of experiments and a performance model, has resulted in a large performance speed-up when compared to the more conventional MC-based implementation. Nevertheless, the simulation times for large AMS circuits are still too large. To solve this problem and to further reduce the simulation times, this section discusses a hierarchical reliability simulator.

First, in Sect. 5.4.1, the problem is discussed in detail. Then, Sect. 5.4.2 proposes a state-of-the-art implementation of the simulator. The simulator not only enables the analysis of large AMS circuits, but also returns a model of the circuit performance including circuit inputs and design parameters. Such a model can be used for fast system-level simulations, but also allows optimization of the circuit design parameters for maximum reliability and performance. Finally, the proposed simulator is demonstrated on an example circuit in Sect. 5.4.3.

5.4.1 Problem Statement

As discussed in Sect. 5.3, the complexity of a circuit increases linearly with the number of devices. Indeed, each device requires a fixed number of factors to characterize the impact of process variability and stochastic aging effects. The simulation

complexity of the DoE-based stochastic reliability simulator also increases nearly linearly with the number of circuit factors (also see Sect. 5.3.3). However, the evaluation time of one circuit sample, which is done with a deterministic reliability simulator, typically increases more than linearly with the circuit complexity. This is related to the inherent need for multiple transient simulations of the entire circuit when performing such a simulation (also see Sect. 5.2). Therefore, stochastic reliability simulation, as proposed in Sect. 5.3, is in practice limited to circuits with around 100 factors or 10–50 devices.

Further, a stochastic reliability simulator requires a user-defined stress bench stating the stress waveforms applied to the circuit input. To evaluate the aging of the circuit for a different stress input, a new and possibly expensive simulation is required. Indeed, although the stochastic simulator generates a circuit model, providing more knowledge about the dominant statistical factors in the circuit and enabling fast time-to-failure simulations, this model only includes statistical circuit parameters.

To solve this problem, a hierarchical reliability simulator is proposed. Such a simulator enables the analysis of large AMS circuits within a reasonable time frame. Additionally, circuit input and design parameters can also be included in the circuit model returned by the simulator. As a result, this model enables fast circuit reliability simulation for different input stress waveforms as well as the optimization of the circuit for maximum reliability and performance.

5.4.2 Implementation

Below, the implementation of the hierarchical reliability simulator is explained. First, the flow of the simulator is overviewed. Then the main components of the simulator are discussed in more detail.

Simulation Flow

Figure 5.23 depicts the flow of the hierarchical reliability simulator. At the input a system description is given. The system consists of one or more netlists and a test bench and a stress bench. System specifications can also be given as input to the simulator. The simulation flow itself is as follows:

1. The system is partitioned in B local subblocks of manageable size (10–30 devices) with only a few terminals (indicated as *Subblock DETECTION* in Fig. 5.23). These subblocks are typically identified manually by the designer according to the hierarchy in the design database (e.g. opamp stages, comparators, filters, etc...), although automatic subblock detection could also be included in the flow. The performance parameters of each subblock are denoted with $\mathbf{P}_b$, with $b = \{1, \ldots, B\}$ and B the subblock number, and are determined by a set of input parameters $\mathbf{u}_b$ and a set of statistical parameters $\mathbf{f}_b$:

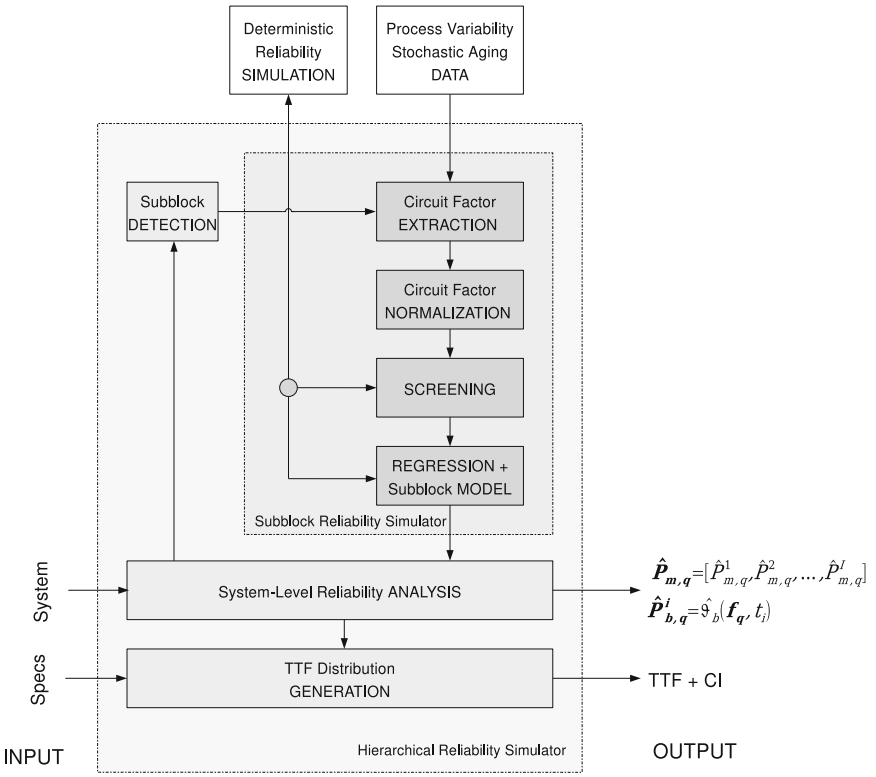


Fig. 5.23 The hierarchical system reliability simulation flow

$$\mathbf{P}_b = f\,(\mathbf{u}_b, \mathbf{f}_b) \tag{5.43}$$

The performance parameters at system level P_m, are then a function of the performance parameters of each subblock:

$$P_m = f\,(\mathbf{P}_b) \tag{5.44}$$

with $m = \{1, \ldots, M\}$ and M the number of system performance parameters. This concept is also depicted in Fig. 5.24 for an example circuit with three unique subblocks (note how subblock two is instantiated multiple times). The performance of each subblock is determined by its input and by a set of statistical parameters. The latter are determined by process variations and stochastic aging effects.

2. Every subblock is modeled separately (indicated as *Subblock Reliability Simulator* in Fig. 5.23). The modeling of each subblock can be done in parallel to reduce simulation time. As demonstrated in the previous section on the stochastic reliability simulator (Sect. 5.3), subblock performance measures (e.g. the input offset voltage, gain-bandwidth, delay, etc.) can be modeled as a weakly nonlinear

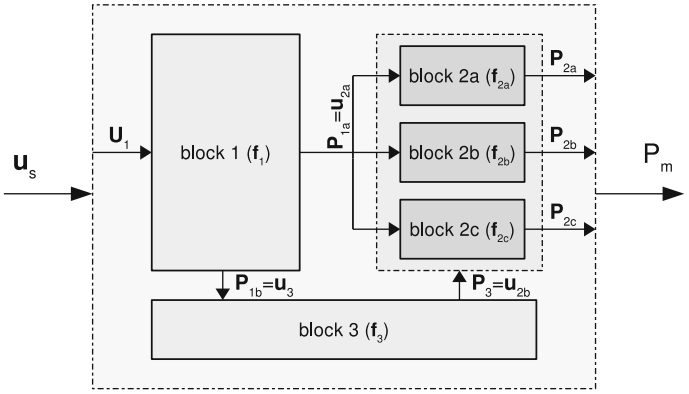


Fig. 5.24 Conceptual representation of a system with subblocks. The performance of each subblock $\mathbf{P}_b, b = \{1, 2, 3\}$, is determined by subblock input parameters $\mathbf{u}_b$ and statistical parameters, $\mathbf{f}_b$ respectively. The statistical parameters are determined by process variations and stochastic aging effects

function of the statistical parameters. The number of statistical parameters is typically around 100 per subblock. To minimize the simulation time, the stochastic reliability simulator defines a well-chosen set of combinations of the statistical parameters (i.e. a DoE), each evaluated with a deterministic reliability simulator. The simulation results then provide the necessary input to build the circuit model. In this approach, each experiment is conducted for a fixed set of input waveforms defined by a circuit stress bench. However, here the subblock input parameters also need to be included in the model, because some subblocks are used multiple times in the same system, albeit each time with a different input (e.g. a comparator in an A/D converter, a NAND gate in a digital circuit, an integrator in a $\Sigma\Delta$-modulator, etc.). Unfortunately, adding these input parameters increases the dimensionality of the problem even further and a more strongly nonlinear behavior can be expected. Both the circuit model as well as the sample selection algorithm (i.e. DoE) therefore need to be altered:

- Sample selection is done with a combination of a *space filling screening design and an active learning sample selection algorithm*. The screening design (indicated as *SCREENING* in Fig. 5.23) consists of a limited number of samples, uniformly distributed over the factor space. New samples are then selected with an active learning sample selection algorithm (indicated as *REGRESSION* in Fig. 5.23). In contrast with the pre-defined samples for the stochastic simulator discussed in Sect. 5.3.3, this sample selection algorithm is adaptive. The algorithm only generates a limited number of samples at a time and uses information on previously taken samples and the current circuit model to generate new samples in areas with the largest uncertainty or model error. New samples are being generated until the model error is sufficiently small or until the pre-defined maximum simulation time is exceeded.

- To model the circuit behavior, a *fast function extraction symbolic regression* is used (indicated as *Subblock MODEL* in Fig. 5.23). Adding circuit inputs as factors for the circuit model possibly results in strongly nonlinear behavior. Therefore, instead of the polynomial model discussed in Sect. 5.3.3, a more appropriate regression algorithm is selected here. This algorithm can handle strong nonlinearities and requires a limited number of samples to be built.

3. The overall system performance is evaluated using the subblock models, instead of actual SPICE-based reliability simulations (indicated as *System Level Reliability ANALYSIS* in Fig. 5.23). This results in a significant simulation speedup and enables the simulation of large analog and mixed-signal circuits.
4. If circuit specifications are provided, the time-to-failure (TTF) distribution is calculated (indicated as *TTF Distribution GENERATION* in Fig. 5.23).

At the output of the simulator, the user receives a model $\hat{\vartheta}_b$ for the performance of each subblock as a function of the subblock's input parameters and statistical parameters and the stress time. Further, the overall system performance $\mathbf{P}_m$, as a function of time, is also returned. If circuit specifications are given, an estimation for the circuit's time-to-failure distribution is provided. Below, more details on the implementation of the regression algorithm and the sample selection algorithm are given.

Multivariate Nonlinear Regression Model

Introducing input parameters in the circuit model, potentially results in strongly nonlinear behavior. The polynomial model used in the stochastic reliability simulator may therefore no longer be adequate and an alternative multi-dimensional regression approach may be needed. The following regression algorithms were considered:

1. multivariate adaptive regression splines (MARS) (Friedman 1991)
2. least-squares support vector machines (SVM) (Suykens and Vandewalle 1999)
3. fast function extraction (FFX) symbolic regression (SR) (McConaghy 2011)

Interpolation algorithms are not considered due to their poor extrapolation performance for high-dimensional problems. A comparison between these multi-dimensional regression techniques has been presented in (McConaghy 2011). Here, it has been shown that the evolutionary-based symbolic regression CAFFEINE (McConaghy and Gielen 2009) and modern feedforward neural networks (Ampazis and Perantonis 2002) are less suitable regression methods for high-dimensional test cases due to the unreasonably long building times or because they are too inaccurate (i.e. test error $> 100\,\%$).

The performance of the remaining regression techniques (MARS, SVM and FFX) is compared here for the 2-dimensional test case shown in Fig. 5.25(left). In this comparison, samples are progressively added to the known set of samples and a model is built for 75% training samples and 25% test samples. The prediction ability

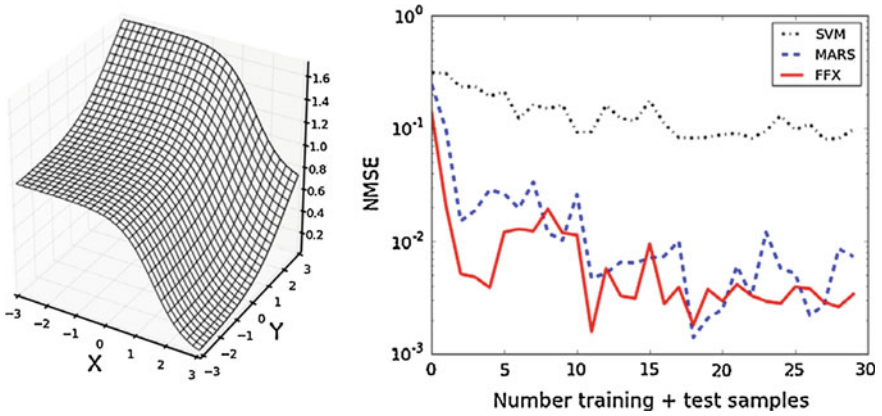


Fig. 5.25 Training and test error (NMSE) for SVM, MARS and FFX for a progressively increasing amount of data points, selected by the active learning sampling strategy (*right*). Test function: $\frac{1.0}{1.0+\exp(2.0(x-1.5))}+\frac{1.0}{1.0+\exp(-1.0(y-2.2))}$ (*left*)

of each regressor is tested by plotting the sum of the test and training normalized mean square error (NMSE) at each generation in the right part of Fig. 5.25.

The error for MARS and FFX easily drops below 1 % when more than 10 samples are available, while the error of SVM stays at approximately 10 %. This is mainly due to the internal regularization paths of MARS and FFX. Here, the regression objective is biased toward cross validation and minimization of the error on the test samples, which prevents overfitting of the data and ill-conditioned model parameters. Moreover, unline MARS, FFX generates a Pareto-optimal set of models that trade off model complexity with test error by ramping up the coefficients in the elastic net formulation (McConaghy 2011). To avoid overfitting even more, a weighted model evaluation can be used by selecting the weights inversely proportional to the test error. The weighted and normalized model formulation $\hat{\vartheta}_{\mathrm{FFX}}$ for K Pareto-optimal models then becomes:

$$\hat{\vartheta}_{\mathrm{FFX}} = \sum_{k=1}^{K} w_k \cdot \hat{\vartheta}_{\mathrm{FFX,k}}, \tag{5.45}$$

with

$$w_k = \frac{(\mathrm{NMSE}_k)^{-1}}{\sum_{r=1}^{K}(\mathrm{NMSE}_r)^{-1}}. \tag{5.46}$$

Because of the ability to pick from a Pareto-optimal set of models, the experiments done further in this work are implemented using the FFX regression algorithm. However, a straightforward extension can be made by building multiple regression methods of a different class simultaneously and to vote or average between them.

Active Learning Sample Selection

The expensive simulation times and high dimensionality of reliability simulations bring about a sparse data set. A full exploration of the parameter space requires the selection of the next sample that is added to the data set to be chosen in such a way that the density of the samples is uniformly distributed. This is the philosophy behind space filling sampling algorithms such as LHS, uniform random sampling, FF designs, etc. (Montgomery 2008). Progressive sampling strategies such as Monte-Carlo random sampling do not necessarily consider previously generated samples and thereby ignore any knowledge about the global sample density and the correlation between the samples. In addition, strongly nonlinear behavior is expected for parametrized signal inputs in the circuit stress bench. Sharp transitions or steep edges in the performance space are preferably sampled more densely than flat or weakly nonlinear regions. Active learning or co-evolution is a supervised machine learning technique where the selection of new inputs is controlled such that the added value of newly gathered information is optimal (Lipson and Bongard 2004). In statistics literature this is described as optimal experimental design (Settles 2009).

The basic setup of active learning sample selection is to predict, for every new iteration of the algorithm, at which locations in the input parameter space one would expect the model to have the highest uncertainty. The uncertainty predictor $D(.)$ is estimated by a distance metric. In this work, the distance metric between two points is expressed as the Euclidian distance or 2-norm $\| \, . \, \|_2$.

First, a distance measure for the input space is declared as follows. Consider $\mathcal{F}_Q \subset \mathcal{F}$ as the collection of Q N-dimensional data samples of the known data set, i.e. the points that already have been simulated. The distance of a newly selected point $\mathbf{f}_{Q+1} \in \mathcal{F}$ to the nearest point $\mathbf{f}_q$, $q = \{1, \ldots, Q\}$, in the known data set $\mathcal{F}_Q$ is then expressed as:

$$D_{\mathcal{F}}\left(\mathbf{f}_{Q+1}\right) = \min_q \|\mathbf{f}_{Q+1} - \mathbf{f}_q\|_2. \tag{5.47}$$

Taking the minimum distance to the known samples forces the algorithm only to look at the nearest neighbor $\mathbf{f}_{qn} \in \mathcal{F}_Q$. A typical space filling sample selection algorithm maximizes the distance function $D_{\mathcal{F}}\left(\mathbf{f}_{Q+1}\right)$ such that newly selected samples are chosen as far as possible from previously visited places. When considering only the input space, the generated model will have the largest uncertainty in these points. Additional information about the nonlinearity of the unknown function and the nature of the regression algorithm itself is accounted for by defining two additional distance functions $D_{\mathcal{P}}$ and $D_{\mathrm{var}(\hat{\vartheta})}$ to be introduced next.

Abrupt changes in the Q known circuit performance parameter values $\mathcal{P}_Q \subset \mathcal{P}$ are predicted by the relative distance of the circuit model output $\hat{\mathbf{P}}_{Q+1} = \hat{\vartheta}(\mathbf{f}_{Q+1})$ to the known output of the nearest neighbor in the parameter space $\mathbf{P}_{q\mathrm{n}} = \hat{\vartheta}(\mathbf{f}_{q\mathrm{n}}) \in \mathcal{P}_Q$. Different circuit parameters are combined into a single distance measure by taking the mean value over all computed relative output distance measures:

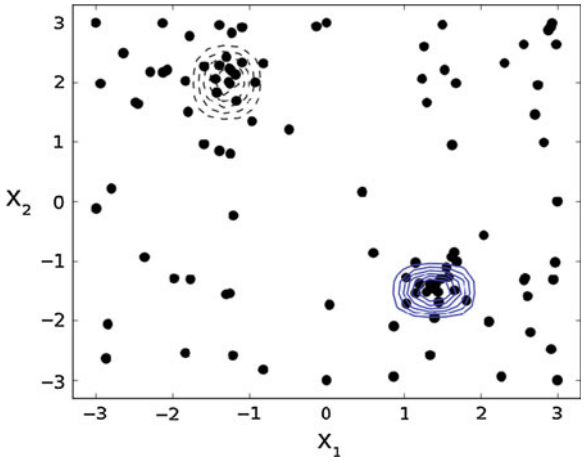


Fig. 5.26 Output distance sampling for the test functions: $f_1(\mathbf{x}) = 4\text{e}5 \cdot \text{sinc}(1.7(x_1 - 1.4)) \times \text{sinc}(2(x_2 + 1.5))$; $f_2(\mathbf{x}) = 3\text{e-}4 \times \text{sinc}(2(x_1 + 1.3)) \times \text{sinc}(1.4(x_2 - 2.0))$

$$D_{\mathcal{P}}(\mathbf{f}_{Q+1}) = \frac{1}{M} \sum_{m=1}^{M} \left(\left\| \frac{P_{m,q\text{n}} - \hat{P}_{m,Q+1}}{\max(\mathcal{P}_Q)_m - \min(\mathcal{P}_Q)_m} \right\|_2 \right). \tag{5.48}$$

with M the dimensionality of $\mathcal{P}$: i.e. the number of circuit parameters. When the distance between the model output and the nearest known output is large, steep edges tend to occur. Adding samples at those locations refines the model by extracting more information at those places. An example of output distance active learning is illustrated on two shifted 2-dimensional sinc functions in Fig. 5.26. It can be seen that more samples are inserted where the peaks (indicated by the contours in Fig. 5.26) occur. Also, optimal samples are taken, considering the behavior of both functions at the same time. This illustrates how the proposed adaptive sample selection algorithm will perform when modeling multiple circuit performance parameters.

A third predictor estimates the model variance and is computed by means of bootstrapping (Hastie et al. 2005). Bootstrapping is a computational method for assessing the model uncertainty. Several deviation models $\hat{\vartheta}_k, k = \{1, \ldots, K\}$, are built for different random permutations of training and test samples. A point in the parameter space where a large variance between the models occurs, corresponds to a large disagreement between the models. This is illustrated in Fig. 5.27. The model variance is normalized to the variance of the median of all deviation models:

$$D_{\text{var}(\hat{\vartheta})}(\mathbf{f}_{Q+1}) = \frac{\sigma^2(\hat{\vartheta}_k(\mathbf{f}_{Q+1}))}{\sigma^2\left[\mu_{1/2}(\hat{\vartheta}_k)\right]} \tag{5.49}$$

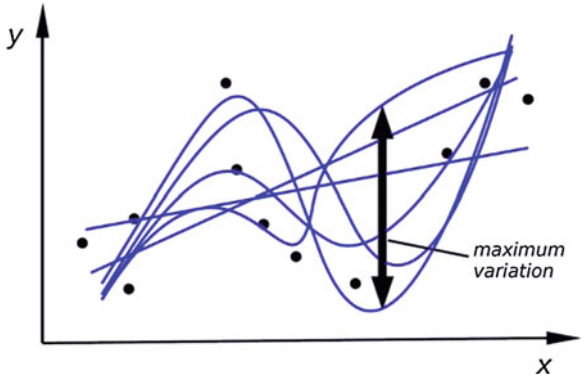


Fig. 5.27 Six bootstrap models for a 2-dimensional example data set. The region of maximal variation between the models is also indicated

with $\hat{\vartheta}(\mathbf{f}_{Q+1})$ deviation model k, evaluated in an unknown point $\mathbf{f}_{Q+1}$ and $\mu_{1/2}(\hat{\vartheta}_k)$ an estimate for the median value returned by all deviation models, evaluated over the entire input space.
Finally, the total distance function used is a combination of the three distance functions (5.47), (5.48) and (5.49):

$$D_{\text{tot}}(\mathbf{f}_{Q+1}) = D_{\mathcal{F}}(\mathbf{f}_{Q+1}) \cdot \left[1 + D_{\mathcal{P}}(\mathbf{f}_{Q+1})\right]^{\alpha} \cdot \left[1 + D_{\text{var}(\hat{\vartheta})}(\mathbf{f}_{Q+1})\right]^{\beta} \tag{5.50}$$

The exponent parameters α and β skew the weight of the distance function towards exploration ($\alpha = \beta = 0$) or towards nonlinearity sampling ($\alpha = \beta = 1$). Note how the total distance function is forced to reach a minimum value when the input distance function equals zero (i.e. when the sample is already included in the known data set).

The next best sample $\mathbf{f}_{Q+1}$, given a known data set $\mathcal{F}_Q$, corresponds to the point where the distance function reaches a maximum:

$$arg \max_{\mathbf{f}_{Q+1}} D_{\text{tot}}(\mathbf{f}_{Q+1}) \tag{5.51}$$

This optimum can be found with a common-purpose global optimization engine such as the multi-objective evolutionary algorithm (MOEA) or with Simulated Annealing approaches (Deb 2001; Kirkpatrick et al. 1983). An example of sample selection using the total distance function D_{tot} is plotted in Fig. 5.28. It can be seen that the sample density is optimally distributed in space and that more dense sampling is encountered at the function peaks. The proposed active learning sample selection algorithm is also compared to random Monte-Carlo-sampling for the test function depicted in Fig. 5.25. Figure 5.29 shows the NMSE of the FFX regression algorithm after each generation using both sampling strategies. On average, the NMSE using

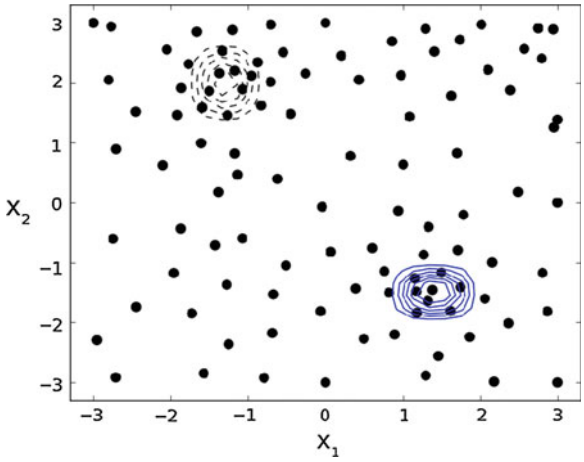


Fig. 5.28 Active learning sample selection using the total distance function $D_{\text{tot}}(\mathbf{f}_{Q+1})$ applied to the test functions depicted in Fig. 5.26

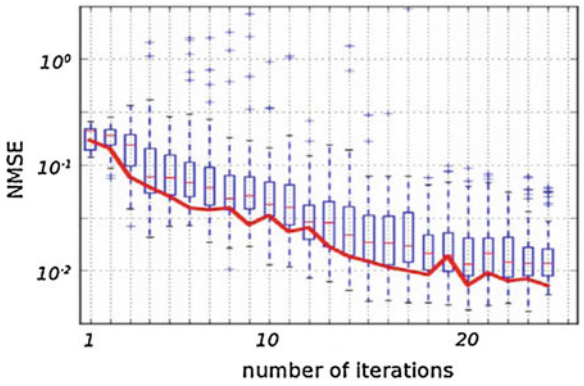


Fig. 5.29 Active learning sample selection versus Monte Carlo for the test case of Fig. 5.25 using FFX models, as a function of the sampling iteration

the active learning selection (indicated with a solid line) drops faster than the 75 percentile of the random sample selection (indicated with a box plot). This demonstrates that, of all possible samples, the proposed algorithm repeatedly selects one of the best samples to further reduce the model error.

The proposed active learning sample selection strategy finds the next best sample, based on information about samples that are already present in the known data set. To start the model building algorithm, an initial data set therefore is needed. To build this data set a space filling DoE such as LHS can be used. Note how the size of this initial data set does not have to increase with the number of dimensions N. It only needs to contain sufficient samples to get the active learning algorithm started.

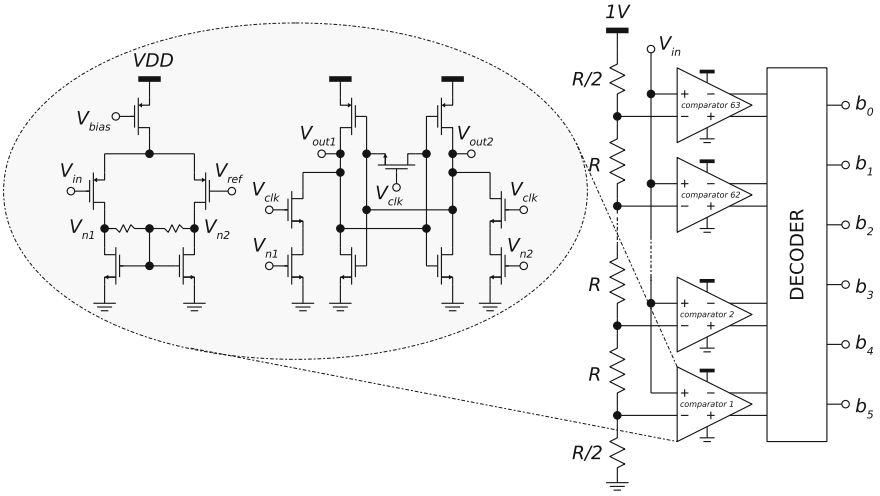


Fig. 5.30 Schematic representation of the demonstrator 6-bit flash ADC. The ADC uses clocked comparators to compare the reference voltages with the input signal

5.4.3 Circuit Example

In this section the clocked comparator, used as a demonstrator circuit in Sect. 5.2 and 5.3, is integrated in a 6-bit flash analog-to-digital converter (ADC). Below, in a first section the converter architecture is explained. Then, the converter is simulated with the hierarchical reliability simulator proposed in this section.

Circuit Schematic

The 6-bit flash ADC test circuit, depicted in Fig. 5.30, is designed in a predictive 32nm CMOS technology with a 1V supply voltage (Arizona state university predictive technology model 2012). The analog part of the circuit consists of more than 1000 circuit elements. The ADC contains 63 clocked comparators, each comparing the input voltage to a different reference voltage. The accuracy of an ADC is typically described by the effective number of bits (ENOB), which is in turn determined by the integral and differential nonlinearity of the converter (INL and DNL respectively (Van De Plasche 1994)):

$$\text{ENOB} = \log_2 \left[\frac{V_{\text{in,min}} - V_{\text{in,max}}}{\max(2 \cdot \text{INL,DNL})} \right] \tag{5.52}$$

Both the INL and DNL are mainly determined by the mismatch between the resistors of the reference ladder and by the input-referred offset of each comparator. Right after

production, both are only determined by process variations. Mismatch, however, can change over time due to BTI effects, as discussed earlier in this chapter.

Simulation Results

The clocked comparator is modeled as a one-system subblock, with the reference voltage as a circuit input that can vary between the ground and the supply voltage. As an input to the ADC, a sine wave with a fixed frequency of 100 Hz, an amplitude of 0.4 V and a DC bias of 0.5 V was applied. Evaluation of the comparator model, build using the active sample selection algorithm and the FFX regression method discussed above, returns a tuple of time-dependent input-referred offset voltages after different stress times.

Figure 5.31 depicts the input-referred offset for each comparator after 1 year of stress and for 100 Monte-Carlo samples, all derived from the comparator subblock model. Comparators at the top and the bottom of the reference ladder are particularly sensitive to transistor aging since they suffer from large asymmetric voltage stress. The bottom comparator for example (i.e. comparator 1 in Fig. 5.30), is at one side stressed by a very low reference voltage, while the other side sees the ADC input (i.e. the sine wave signal). Since NBTI is exponentially dependent on the magnitude of the gate voltage stress, this results in a large threshold voltage mismatch between the input transistors (on average $\Delta V_{\mathrm{TH}} = 17\,\mathrm{mV}$ after 1 year for comparator 1). A similar effect can be observed for comparators at the top of the reference ladder (e.g. comparator 63 in Fig. 5.31). The input offset increases over time and results in a reduction of the ENOB. Figure 5.32 shows a normal probability plot of the ENOB right after production, after 1 month of operation and after 1 year. The solid lines are the ENOB as computed by the hierarchical models of the comparators, while the markers represent the ENOB calculated from a full system aging simulation with a MC-based stochastic aging simulator as explained in Sect. 5.3.[10] From Fig. 5.32, it can be seen how process variations cause a large initial spread on the ENOB, while the graph shifts towards lower values due to aging effects.[11] It is clear how the results of the hierarchical simulation method presented here (i.e. solid lines) agree very well with the results predicted by the stochastic reliability simulator which is not using models for each system subblock. A good correspondence between the model and the full system simulations is observed. Moreover, the logarithmic time dependence

[10] Given the large amount of circuit factors in the ADC (>1000), building a model with the DoE-based stochastic reliability simulator is computationally too expensive (also see Fig. 5.17). Therefore, as a benchmark the MC-based stochastic reliability simulator is used and evaluated in only ten samples.

[11] On average, the ENOB shifts with almost one bit over a time span of one year. The reason for the large shift is the lack of offset compensation in the comparators, the application of a fixed stress pattern at the input of the circuit and the use of an aging-sensitive 32 nm CMOS technology. In reality, the expected degradation will possibly be smaller. Nevertheless, the example indicates how aging could potentially be a large problem and how the presented simulation technique can be used to analyze these cases.

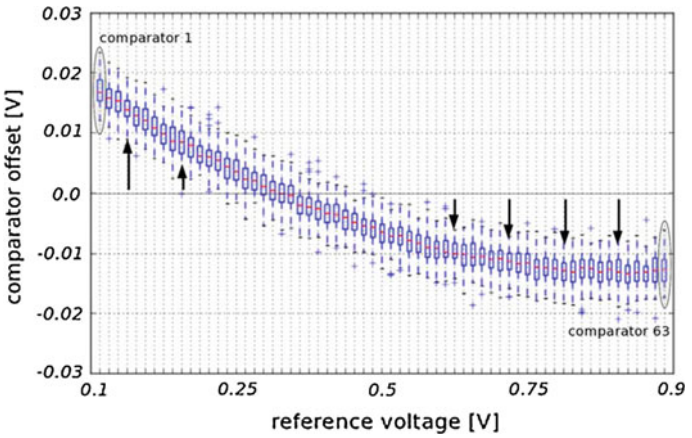


Fig. 5.31 The input-referred offset voltage for each flash ADC comparator after 1 year of stress and for 100 random samples, all evaluated with the comparator subblock model. The *arrows* indicate the offset shift between the initial state and after wearout

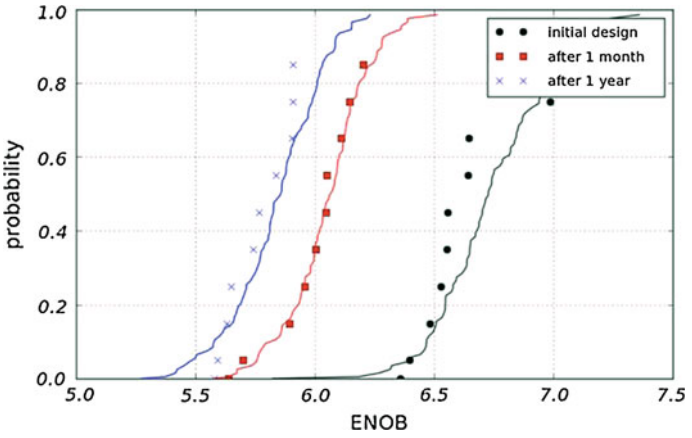


Fig. 5.32 A normal probability plot of the effective number of bits of the flash ADC for 100 random samples evaluated with the proposed hierarchical simulator (*solid lines*) and 10 samples evaluated with full system aging simulations (*markers*). The ENOB represents the static accuracy of the converter and decreases over time due to transistor aging effects

of the NBTI effect is also observed. The discrepancy at time zero is due to the limited number of full system simulations (only ten).

The demonstrator circuit has been simulated on a dual-quad core 2.8 GHz Intel Xeon processor with 8 GB of RAM. Model build time for the comparator subblock took 31 min, while evaluation of the entire converter took 1 min and 41 s for 100 Monte-Carlo samples. Evaluation of just one Monte-Carlo sample, using a deterministic

reliability simulator (see Sect. 5.2) took 1 h and 55 min. This results in a speedup of 360× when evaluating 100 random samples with both methods.

5.5 Conclusions

As discussed in Chap. 2, transistor aging is an emerging problem, especially for circuits integrated in nm CMOS processes. This problem needs to be addressed at design time. Therefore, transistor compact models and simulation techniques to accurately estimate the impact of aging on a specific design are required. Accurate transistor aging compact models have been presented in Chap. 3. Existing academic and commercial reliability simulators have been discussed in Chap. 4. The latter, however, still have many deficiencies.

In this chapter, each of the issues listed in Sect. 4.4 has been addressed. Three reliability simulators have been presented. The first method, a deterministic reliability simulator, focuses on deterministic aging effects. The proposed method enables fast and accurate reliability simulation, includes the combined impact of multiple aging effects and enables a sensitivity analysis to detect circuit reliability weak spots. Next, a stochastic reliability simulator has been proposed. This simulator includes the impact of process variations and stochastic aging effects. The latter are especially important in nm CMOS, when even BTI and HCI effects become stochastic. Furthermore, the simulator enables an extensive analysis of the circuit at hand yielding time-dependent circuit performance, time-to-failure analysis, capability to evaluate a degraded netlist in SPICE, etc. In addition, the method also provides a model of the circuit performance as a function of the most important statistical parameters. Finally, a hierarchical reliability simulator has been presented. This simulator allows the simulation of large analog circuits in a limited time frame but with the accuracy and capabilities provided by the stochastic reliability simulator.

Each of the three simulators has been demonstrated on an example circuit. First, the average degradation of a clocked comparator has been analyzed with the deterministic reliability simulator. Then, the impact of stochastic effects on the same comparator circuit has been studied with the stochastic reliability simulator. Finally, the comparator has been integrated as part of a 6-bit flash ADC and the entire system has been analyzed using the hierarchical reliability simulator. Where appropriate, the simulation results have also been verified with hand calculations using a first-order BTI model.

The set of simulation methods proposed in this chapter, now allow to easily analyze any analog circuit. This will be done in the next chapter, where the impact of transistor aging on typical analog circuits will be studied. The result of this study will provide a better understanding of which analog circuits are most sensitive to aging.

Chapter 6
Integrated Circuit Reliability

6.1 Introduction

IC non-idealities such as process variations and transistor aging are a potential threat for the correct operation of integrated circuits. Furthermore, aging effects become more important for circuits integrated in sub-45 nm technologies (see Chap. 2). To guarantee circuit reliability over a product's lifetime, a foundry typically performs accelerated stress measurements on individual devices and calculates the maximum transistor operating voltages allowed in the technology process. The latter are typically defined as the stress voltage for which the drain current or threshold voltage does not exceed a given reliability margin (e.g. $\Delta I_{DS}/I_{DS} < 10\,\%$ or $\Delta V_{TH} < 50\,\text{mV}$ after 10 years) (also see Sect. 1.5). This does however not guarantee reliability of a circuit since:

- The reliability margin is chosen arbitrarily and the sensitivity of the circuit to individual transistor variations is not considered.
- Aging-induced transistor variations due to stochastic aging effects are not taken into account.

Better performing and guaranteed reliable circuits can be realized if the actual impact of the degradation mechanisms on the circuit performance is evaluated during the design phase. This chapter aims to provide a designer with a better understanding of how transistor aging can affect the performance of a circuit. First, the focus is on the different aspects that determine the lifetime of a circuit; then a reliability-aware design flow is demonstrated on an example circuit. The outline of the chapter is as follows. Section 6.2 studies the impact of transistor aging on the operation of typical analog building blocks. To do this, the transistor aging models proposed in Chap. 3, and the circuit simulation methods developed in Chap. 5, are used. The information extracted from this assessment, can be adopted by designers to anticipate potential reliability problems, to alleviate them or to avoid unnecessary overdesign. Next, Sect. 6.4 demonstrates the latter on a fully elaborated example current-based DAC circuit. The DAC circuit is first designed using a conventional design flow

E. Maricau and G. Gielen, *Analog IC Reliability in Nanometer CMOS*,
Analog Circuits and Signal Processing, DOI: 10.1007/978-1-4614-6163-0_6,

with device-based guardbanding. Then the power-area product, which is a metric for the amount of resources needed to build and operate the circuit, is minimized using a reliability-aware design flow combined with circuit design techniques. In Sect. 6.5, the transistor models and simulation methods developed in this work are demonstrated on a digital datapath circuit. Although the focus of this work is on analog circuits, it is also possible to analyze small- to medium-sized digital blocks such as a standard cell, a datapath or an SRAM cell. Finally, chapter conclusions are given in Sect. 6.6.

6.2 Assessment

This section studies the factors that determine the lifetime of an analog integrated circuit. The lifetime is defined as the time to first circuit failure or the time to reach a point where the circuit fails to meet its application-dependent specifications such as minimum gain or maximum offset.[1] All relevant transistor aging effects are included in the simulations and the circuit is assumed to operate at nominal conditions (i.e. circuit failure due to transient effects such as EMI or supply noise is not considered). The aim of this section is not to provide an extensive and complete overview of the dependencies of each analog circuit, but rather to indicate the important factors to look at when designing for reliability. The concepts explained and illustrated in this section are then applied to an example circuit in the next section.

The time-dependent degradation of each transistor in a circuit is a function of the technology and the circuit stress conditions, topology and sizing (see Fig. 6.1). The aging-induced transistor performance shift, combined with the observed performance parameter and the process capability index, determines the circuit lifetime. Each of these factors is discussed in more detail and illustrated with examples in the sections below. First, Sect. 6.2.1 explores the relationship between different circuit performance parameters and the sensitivity to circuit aging. Then, the link between the process capability index or C_{pk}, which expresses the amount of design guardbanding, and the sensitivity of a circuit to failure is reviewed in Sect. 6.2.2. Section 6.2.3 discusses the impact of technology on transistor aging and Sect. 6.2.4 assesses the impact of the circuit design itself. Finally, Sect. 6.2.5 studies the influence of the circuit stress conditions. The latter includes the waveforms applied to the circuit input, the operating voltage, the environmental temperature and the stress time.

[1] These limits are set by the application engineer. The performance of most analog circuits shifts gradually and result in a parametric failure, a complete functional failure is very unlikely. Digital circuits on the other hand are typically clocked. Therefore, a gradual shift of the transistor characteristics is not always visible until timing criteria are no longer met. At that point the circuit performance specifications are no longer met and a complete functional failure is very likely.

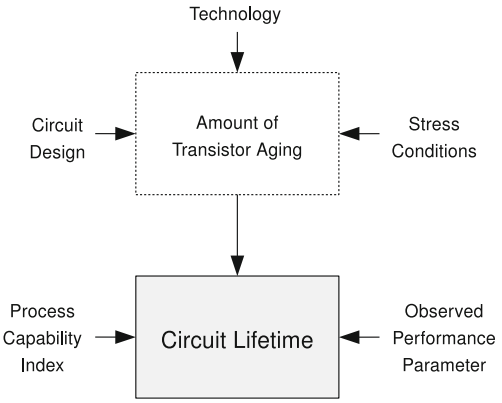


Fig. 6.1 The lifetime of an integrated circuit is a function of the observed performance parameters, the process capability index and the amount of transistor aging. In turn, the latter depends on the stress conditions, the technology and the circuit topology and sizing

6.2.1 Observed Performance Parameter

The correct operation of a circuit is typically expressed with a number of performance parameters. For digital circuits, delay and power consumption are the most important parameters and a standard cell design becomes unreliable if the delay increases too much and it results in a timing violation. Quantifying the performance of an analog circuit, however, is more complex. For an amplifier for example the gain, the gain-bandwidth, the offset, the power consumption, the phase margin and the common-mode rejection ratio are a subset of possibly crucial circuit parameters. The relevant circuit parameters depend on the application. Therefore, when evaluating the circuit reliability, it is important to first identify these relevant performance parameters. Each of these parameters must then meet its specification over the entire intended circuit lifetime.

To obtain a better understanding of which circuits are sensitive to failure, six commonly used analog circuits are examined: an LC-VCO, a comparator circuit, a symmetrical OTA, a current mirror, a one-stage resistive amplifier and a MOS resistor. Each circuit is designed in both a 65 nm and a (predictive) 32 nm CMOS process and is tuned such that the performance at time zero is identical in both technologies. The circuits are evaluated over a lifetime of one year using the simulation framework described in Chap. 5. Table 6.1 depicts the results of this analysis. For each circuit, different circuit performance parameters P were monitored. Both the average performance shift $\Delta\mu(P)$ and the shift in standard deviation $\Delta\sigma(P)$ were recorded over time. In Table 6.1, critical values are indicated in bold.

It is clear that, for a given circuit, particular performance parameters change a lot, while other parameters don't change at all. For example, the oscillation amplitude of the LC-VCO reduces with 16–18 %, while the oscillation frequency does not change

Table 6.1 Reliability analysis of analog circuits for a 65 nm and (predictive) 32 nm CMOS process with a stress time of one year

			65 nm CMOS		32 nm CMOS	
	Performance P	Nominal value	$\Delta\mu(P)(\%)$	$\Delta\sigma(P)(\%)$	$\Delta\mu(P)(\%)$	$\Delta\sigma(P)(\%)$
LC-VCO	Frequency	2.8 GHz	0.0	0.0	0.0	0.0
	Amplitude	0.7 V	**−18**	**+62**	**−16.4**	**+59**
	Phase noise @1MHz	−109.4 dBc/Hz	**−8**	+2.0	**−4**	+1.2
Comparator	Offset	1e-6 V	**+1800**	+0.2	**+13900**	+0.22
	Slew rate	3.2e8 V/s	−0.1	+0.5	−0.9	−1.0
Symmetrical	Gain	35 dB	0.0	+0.9	+0.03	+3.0
OTA	Bandwidth	34 MHz	0.0	0.6	−0.5	+2.0
Current mirror	I_{out}	8μA	0.0	0.0	**−0.6**	**+6.8**
	ΔI	4.5 %	0.0	+0.6	**+17**	**+6.6**
R-amplifier	Gain	10 dB	0.0	0.0	−0.5	+4.4
	Bandwidth	8 MHz	0.0	0.0	0.0	+0.7
MOS resistor	R_{on}	400 Ω	+0.1	0.0	**+6.5**	**+38**

Critical values are indicated in bold

at all. Indeed, the oscillation frequency of this oscillator largely depends on the value of the oscillator tank inductor and capacitor. These passive components do not age.[2]

The oscillation amplitude, however, changes over time since it depends on the gain of a cross-coupled transistor pair. This gain reduces over time due to HCI and PBTI effects. Below, the circuit performance parameters that are immune to transistor aging and the aging-sensitive performance parameters are discussed.

Aging-immune Circuit Performance Parameters

Circuit parameters that are insensitive to transistor process dependent parameter variations are typically also immune to transistor aging. Indeed, when subjected to nominal operating voltages, aging-induced transistor parameter shifts are rather small (i.e. $\Delta V_{TH} < 30\,\text{mV}@1\text{year}$) and have little impact on circuit parameters such as:

- The oscillation frequency of an LC-VCO (also see Table 6.1).
- The input-output behavior of a circuit with a passive feedback network (e.g. an active-RC filter).
- The gain, slew rate and bandwidth of an amplifier (also see Table 6.1).

Large or additional guardbands, to guarantee reliable circuit operation over the lifetime, is thus not necessary if the circuit performance is determined by these aging-immune parameters.

[2] As discussed in Sect. 2.4, electromigration can affect the performance of passive components, but this is not taken into account here.

Aging-sensitive Circuit Performance Parameters

From the results shown in Table 6.1, three causes for significant and possibly problematic circuit performance shifts, and associated circuit reliability problems, can be identified:

1. When subjected to *large stress voltages* (i.e. $V_{GS,max}$, $V_{DS,max} \gg V_{DD,nom}$), transistor parameters can shift significantly within the intended operational lifetime of a circuit ($\Delta V_{TH} > 0.1$V after 1 year). In that case, even performance parameters of analog circuits that are not sensitive to small transistor parameter variations can induce a reliability problem. The oscillation amplitude of the LC-VCO in Table 6.1, for example, reduces significantly due to large voltages applied to the cross-coupled transistor pair and the resulting HCI-induced degradation. In turn, this is reflected in the phase noise, which reduces over time (also see Fig. 6.2 showing the phase noise degradation over 5 years).
2. Asymmetrical stress, applied to a symmetrical circuit, can result in time-dependent mismatch. This effect is demonstrated in Fig. 6.3 where asymmetric stress at the input of a clocked comparator circuit results in a time-dependent change of the input offset. When used as part of a larger system, this offset degradation may affect circuit parameters such as the linearity of an A/D converter (also see Sect. 5.4).
3. Due to stochastic transistor aging effects in 45 nm and below (see Chap. 3), transistor parameter variations can increase over time. As a result, matched transistors can suffer from time-dependent mismatch, even when they are stressed identically. Examples of circuit performance parameters that are affected by these stochastic aging effects are the output current of a current mirror and the on-resistance of a MOS resistor. Numerical simulation results for both circuits are given in Table 6.1, while Fig. 6.4 depicts the degradation behavior of the MOS resistor in more detail. The slope of the CDF clearly decreases over time, indicating an increase of the standard deviation.

6.2.2 Process Capability Index

The process capability index or C_{pk} is a metric to express the relationship between process and aging induced circuit performance variations and the minimum required circuit performance:

$$C_{pk} = \min\left(C_{pl}, C_{pu}\right) \tag{6.1}$$

$$\text{with} \quad \begin{cases} C_{pl} = \frac{\mu(P) - P_{min}}{3\sigma(P)} \\ C_{pu} = \frac{P_{max} - \mu(P)}{3\sigma(P)} \end{cases}$$

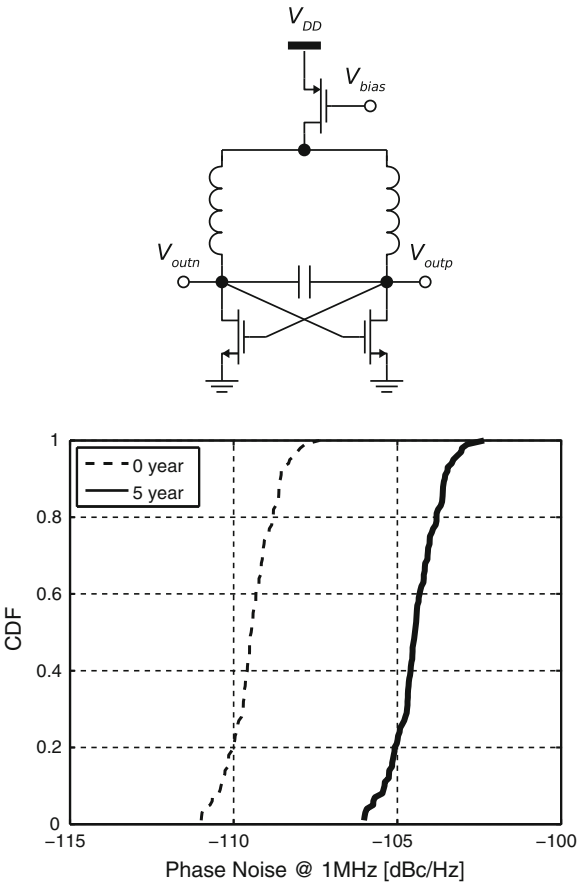


Fig. 6.2 An LC-VCO (*top*) can have a large time-dependent decrease in the phase noise due to HCI and PBTI induced reduction of the oscillation amplitude (*bottom*). The circuit is designed in a predictive 32 nm CMOS technology, has an operating voltage of 1.25 V and an oscillation frequency of 2.8 GHz

with $\mu(P)$ the average circuit performance and $\sigma(P)$ the standard deviation on that average. $P_{\min}$ is the lower specification limit and $P_{\max}$ the upper specification limit on performance P. A large C_{pk} value implies a high yield. Typically, a C_{pk} larger than one is required, since this corresponds to a yield of 99.7 %. However, obtaining a large C_{pk} typically also requires more area and a larger power consumption. Due to transistor aging, both the mean performance $\mu(P)$ and the spread on the performance $\sigma(P)$ can shift over time. This will result in circuit failure for samples crossing the upper or lower specification limit. Circuits with a low initial C_{pk} are thus more prone to circuit aging.

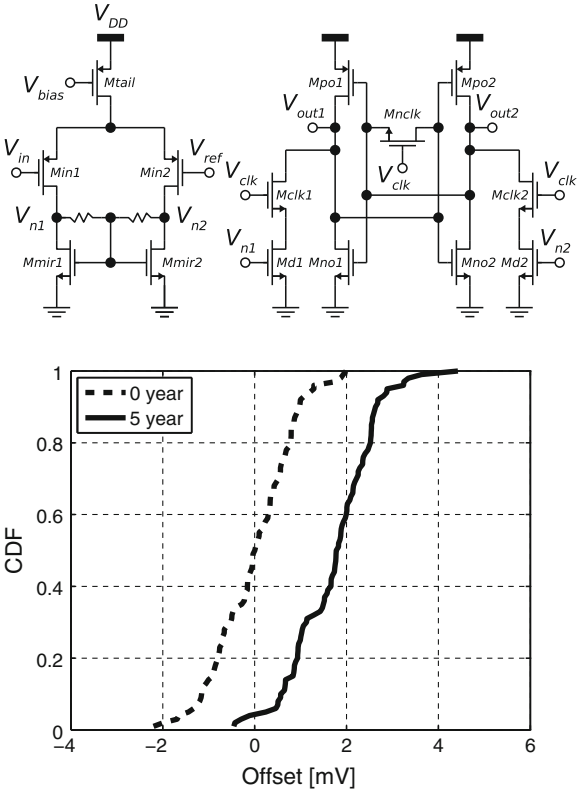


Fig. 6.3 A comparator (*top*) can suffer from asymmetrical stress resulting in a time-dependent shift of the input offset due to NBTI effects in the input transistor pair (*bottom*). The circuit is designed in a predictive 32 nm CMOS technology and has an operating voltage of 1.0 V. At one input a reference voltage of 0.2 V is applied, the other input is stressed with a 0.4 V amplitude sine wave around a 0.5 V DC bias

The relationship between the C_{pk} and the circuit lifetime is illustrated with an example in Fig. 6.5. The time-dependent performance shift of two MOS resistor circuits is observed. Both circuits have the same initial average resistance. One MOS resistor, however, is sized three times larger than the other and is less sensitive to process variations. Circuits which have a resistance that deviates more than ten percent from the average value are denoted as a failure. The resulting initial C_{pk} value of the first circuit is then 1.5, while the second circuit has a C_{pk} of 2.45. After five years of operation, however, the mean and standard deviation of the resistance of both circuits has increased. This results in a time-dependent increase in the amount of circuit failures and a reduction of the C_{pk}. After five years the smallest circuit obtains a C_{pk} of only 0.47; the second circuit, however, still has a $C_{\mathrm{pk}} = 1.0$ due to its more robust, but larger, initial design. The C_{pk} is thus a good metric to express the amount of design guardbanding for a specific circuit performance parameter. In case this

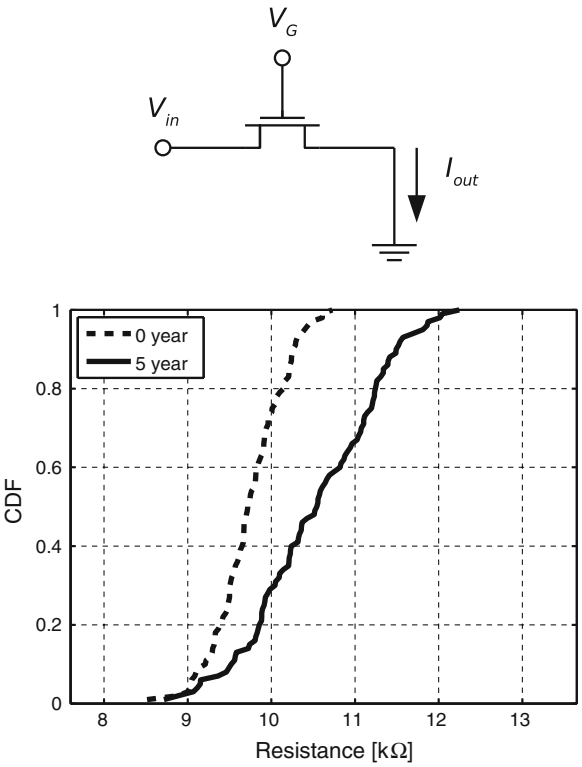


Fig. 6.4 A MOS resistor (*top*) can have a large time-dependent increase in the mean resistance value and in the standard deviation on the resistance due to PBTI effects (*bottom*). The circuit is designed in a predictive 32 nm CMOS technology. A 1.0V gate voltage and a 0.5 V input voltage are applied

parameter is sensitive to aging, designing for a larger initial C_{pk} will result in a larger guardband and a more reliable circuit. Simulation tools such as described in the previous chapters allow a designer to calculate the necessary guardband for a designer.

6.2.3 Technology

Table 6.1 lists the circuit performance shifts for various analog circuits, each integrated in a 65 nm process and a (predictive) 32 nm CMOS process. The 32 nm implementations are clearly more sensitive to transistor degradation. Ever-increasing oxide electric fields, the severely aggravated PBTI effect in nMOS transistors and the increased time-dependent variability due to stochastic aging effects are the main causes for this result (also see Chap. 3). As discussed in Sect. 2.2, for example,

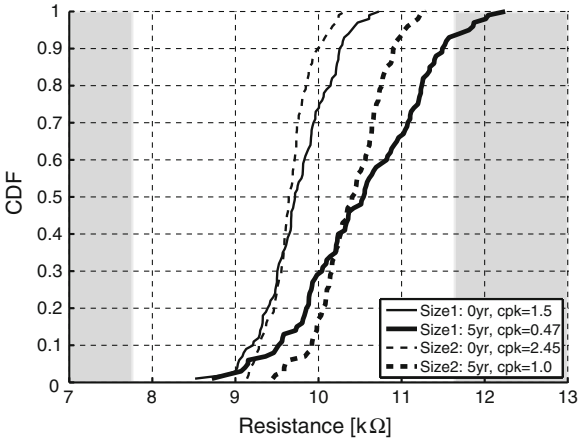


Fig. 6.5 Reliability analysis for two MOS resistor circuits with different sizes (size1$=0.3\times$size2). Both circuits are designed in a predictive 32 nm CMOS technology. For each circuit, the C_{pk} is calculated, right after production and after five years of operation. The upper and lower failure region are indicated in gray on the figure. The larger sized MOS resistor is less sensitive to process variations and stochastic aging effects and therefore retains a C_{pk} larger than one, even after five years of operation

the oxide electric field nearly doubles when going from a 65 to a 32 nm technology. When designing in advanced CMOS technologies, it is therefore important to help a designer to estimate the impact of aging on his or her circuit in order to find a good trade-off between reliability and performance. The transistor compact models discussed in Chap. 3, are also suitable for ultra-scaled technologies. Furthermore, Sect. 3.6 has even proposed a first-order model for hand calculations with model parameters for a 32 nm CMOS technology. This model has been demonstrated in Chap. 5 and enables designers to calculate approximately the average threshold voltage and the variance on the threshold voltage as a function of time.

6.2.4 Circuit Design

The circuit design, determined by the topology and the transistor sizing, also affects the time-dependent transistor aging. How transistor sizing can affect circuit lifetime has been demonstrated in Sect. 6.2.2 were a MOS resistor circuit was shown to be more reliable if the area is larger. The impact of the topology on transistor aging is illustrated in Fig. 6.6. The clocked comparator circuit, used as an example circuit throughout Chap. 5, is compared to a different implementation of the same building block. Both circuits are designed in a (predictive) 32 nm CMOS technology and are subjected to NBTI and PBTI. The difference between the two topologies is the nature (pMOS or nMOS) of the input transistors in the output stage. Both circuits are

subjected to their respective worst-case input stress ('1' at one input, '0' at the other input) and the aging-induced input offset shift is analyzed. The result is depicted in Fig. 6.6. The comparator with a pMOS input stage clearly ages more than the nMOS counterpart. Both circuits age due to BTI effects in the input transistor pair. Due to the stacked current source and the input pair transistor in the pre-amplifier, however, the voltages V_{n1} and V_{n2} never reach the supply voltage. Therefore, in the pMOS implementation, Md1 and Md2 are stressed more than in the nMOS implementation. Eventually, this results in a larger overall degradation and a larger input offset shift of the comparator. This example clearly shows how the worst-case degradation can differ for different implementations of the same building block. As a designer, the information returned by the reliability simulator is therefore crucial to adapt the circuit topology in such a way that it suffers less from aging while always performing according to specifications over the circuit's lifetime.

6.2.5 Stress Conditions

The stress conditions applied to the circuit can have a large impact on the shift of circuit parameters. These stress conditions are determined by the applied circuit input, the operating voltage, the temperature and the stress time. Figure 6.7, for example, demonstrates the impact of various input voltages and reference voltages on the input offset of the same comparator circuit in a 32 nm CMOS process. Both the reference voltage and the input voltage have a large impact on the input offset which, in this case, varies between −3.48 and 4.96 mV depending on the applied input. A reliability analysis of a circuit should therefore also include the impact of the circuit inputs. Searching for realistic worst-case stress conditions is a difficult problem and an important challenge to be solve in the coming years. An efficient reliability simulator able to generate circuit models, as discussed in Sect. 5.4, can be used as an aid for such as study.

6.3 Failure-Resilient Circuits

Section 6.2 has reviewed various parameters that affect the lifetime of an analog circuit. The impact of each of these effects can be simulated with the models and tools discussed in Chaps. 4 and 5. This information can then be used to design guaranteed reliable integrated circuits, being:

1. Intrinsically robust circuits.
2. Self-healing analog circuits which use the huge potential of digital circuits to continuously monitor the operation of the core circuit and to compensate for errors at run time.

Both approaches are briefly discussed in the following sections.

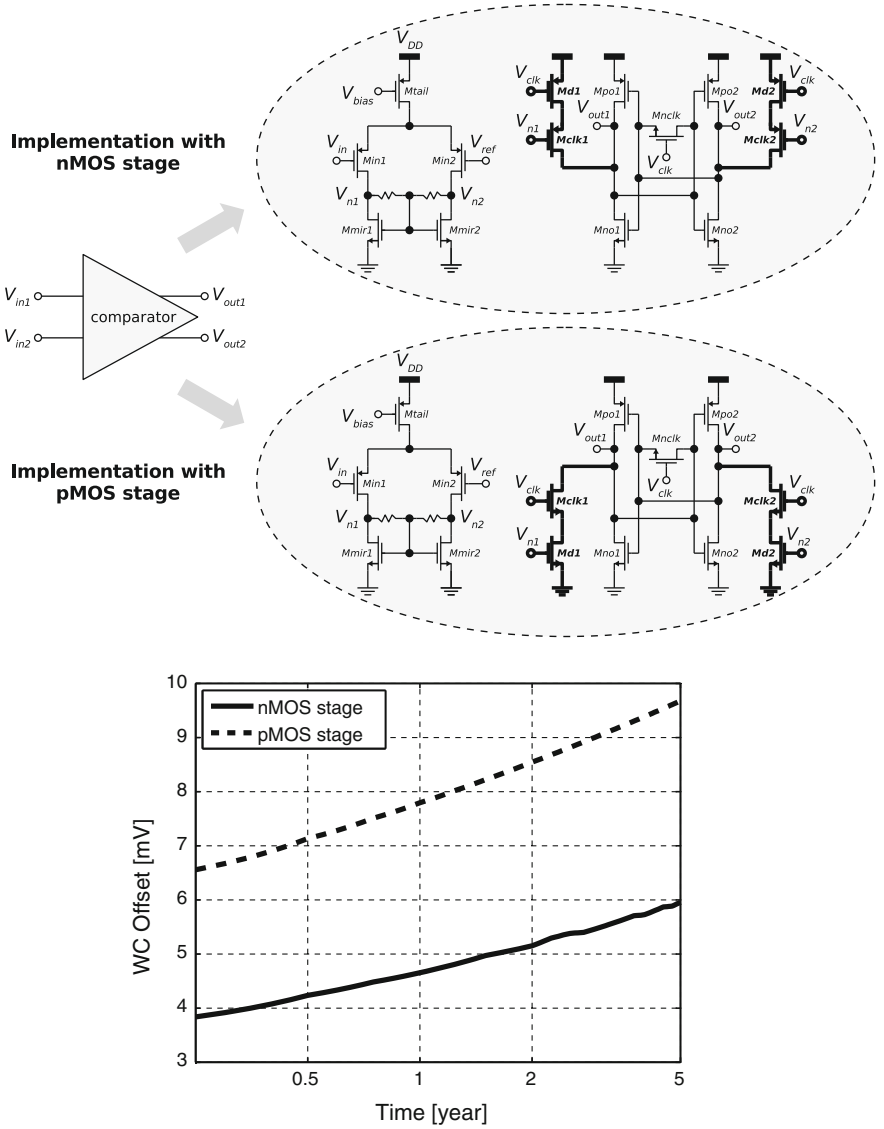


Fig. 6.6 Two implementations of a clocked comparator circuit, implemented in a predictive 32 nm CMOS technology. Both circuit age due to a BTI effect in the input transistor pair but, depending on the nature of the implementation, a different time-dependent offset shift is observed

6.3.1 Intrinsically Robust Circuits

Classically, circuits are designed for intrinsic robustness by overdesign (guardbanding) or the use of redundancy. This results in a large power and area penalty. However,

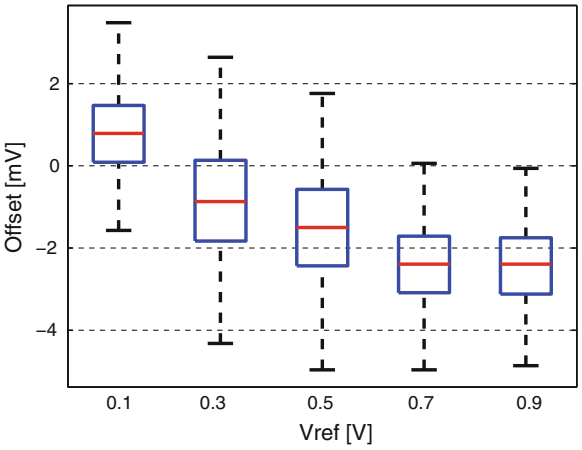


Fig. 6.7 Boxplot of the input offset shift of a comparator circuit for different reference voltages. For each reference voltage, at the comparator input a square wave voltage with a duty factor ranging from 0.0 to 1.0 in steps of 0.25 was applied

an accurate and efficient circuit reliability analysis tool can provide detailed information about the yield and operation of the circuit over its entire lifetime. Therefore, this tool can help to significantly reduce the amount of overdesign, while still creating a robust circuit.

As a simple illustration, the concept is demonstrated on a 90 nm CMOS one-stage resistive amplifier (see Fig. 6.8). Two performance parameters are monitored: the AC output voltage V_{out} and the DC output voltage V_{OUT}. The circuit is simulated over a stress time of four months in which the circuit degrades due to hot carrier effects. The reliability simulator has indicated the transistor length L as a circuit weak spot for aging. A large L improves the robustness of the circuit to aging. Therefore, a second amplifier in which both the width and the length of the transistor are doubled, was designed and simulated. This is expected to improve the circuit reliability, while the circuit performance remains the same and the overall power consumption does not increase. Simulation results for both the original and the improved circuit shown in Fig. 6.8 confirm this. The circuit performance parameters of both circuits shift over time due to hot carrier degradation. However, the second circuit with the larger transistor ages much less than the original circuit and therefore the robustness has been improved significantly.

Shanbhag (Shanbhag et al. 2010) proposed another solution by solving the reliability problem at system level with stochastic computation techniques. Variations at transistor level are viewed as noisy communication channels and communications-inspired design techniques based on estimation and detection theory are proposed to guarantee reliable (i.e. acceptable) circuit operation at all times. Also, Chakradhar (Chakradhar and Raghunathan 2010) suggested a third approach. He proposed to perform

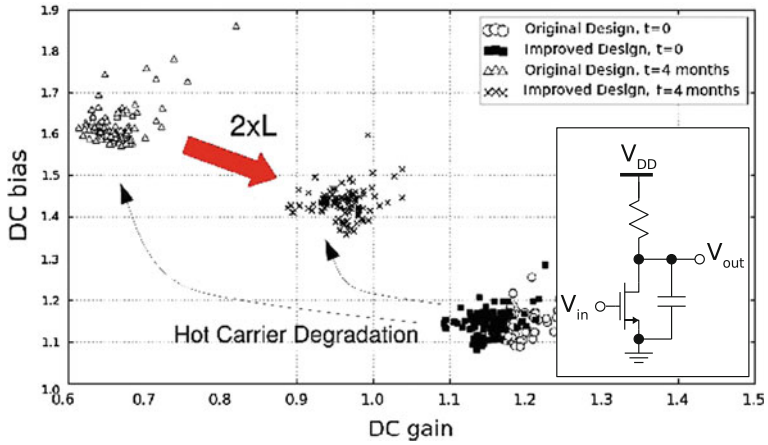


Fig. 6.8 Illustrative reliability simulation of a one-stage amplifier over a lifetime of four months. The circuit has been designed in a 65 nm CMOS technology. The DC and AC output voltage of the circuit shift over time due to hot carrier degradation. An improved circuit (i.e. with larger transistor length) is more resilient to hot carrier degradation and therefore ages less

computations with best effort, but to tolerate errors if they occur. This approach is only possible in non-critical applications such as real-time video.

6.3.2 Self-healing Circuits

Section 6.3.1 advocates the usage of local overdesign, local redundancy or system-level solutions to design an intrinsically robust system. For some stochastic reliability problems, however, this is not an optimal solution. For example, when designing an analog-to-digital converter, each output driver has to be sized such that it can switch fast enough to guarantee the minimum speed required for that converter. Under the presence of process variability, however, this will result in a huge overdesign and extra power consumption for the entire converter to guarantee high yield.

A solution for this problem can be found in the use of the 'knobs and monitors' (also referred to as 'sense and react' (Rabaey et al. 2011)) principle. The idea is to monitor the operation of a system or circuit and take runtime countermeasures to compensate for variability and reliability errors. This guarantees a correct and optimal operation at all times, if properly anticipated for the necessary 'tuning' range at design time by using analysis and design tools. Note how the knobs and monitors principle is similar to the concept of digitally assisted analog circuits (Murmann 2006). The latter tunes out process variability or other nonideality effects in an analog block with the usage of digital peripheral circuits. To also monitor transistor aging effects, however, additional or different positions for the knobs and the monitors will be needed. This therefore requires novel analysis tools and design techniques.

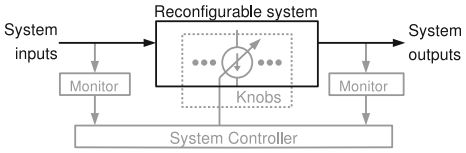


Fig. 6.9 General architecture of a knobs and monitors (sense and react) based system

As shown in Fig. 6.9, a knobs and monitors system consists of 3 parts. Monitors inserted in or added to the circuit measure the actual performance of the circuit at run time with simple measurement circuits. Knobs are tunable or reconfigurable circuit parts able to change the operating point of the system. Finally a Control Algorithm in the controller selects, based on the inputs from the different monitors, the optimal configuration of the system knobs in order to satisfy the system specifications, even if the performance varies over time. The control loop can be implemented in digital hardware, adding only a limited amount of extra power dissipation and area consumption.

Singh et al. (2010); Karl et al. (2008) and keane et al. (2010) proposed circuits for in-situ degradation monitoring of BTI, HCI and TDDB aging effects. De Wit and Gielen (2012) and Acar et al. (2008) applied the knobs and monitors principle on an output driver to guarantee a minimum power efficiency. Figure 6.10 (left) depicts a schematic representation of such a driver (De Wit and Gielen 2012). Analysis in a 90 nm CMOS technology indicates how process variability, NBTI, hot carriers and breakdown effects affect the characteristics of the output stage transistors and reduce the conversion efficiency. To monitor this effect, the on-resistance of the output-stage, which is directly linked to the conversion efficiency, is measured periodically. If the on-resistance increases too much, extra output transistors are switched on in parallel to the already active but aged output transistors. This system guarantees an optimal performance versus power consumption trade-off over the entire circuit lifetime (De Wit and Gielen 2012). Note how circuit monitors are also subjected to aging effects. However, the circuit performance is only measured every few weeks (i.e. depending on how fast the circuit ages). Therefore the monitor circuit degradation is only very limited and wrong decisions due to aged monitors are avoided.

6.4 Case Study 1: IDAC

Here, the transistor models and simulation methods proposed earlier in this work are used to demonstrate the design of a failure-resilient circuit using a reliability-aware design flow. The example circuit is a current-steering digital-to-analog converter (IDAC), implemented in a (predictive) high-k 32 nm CMOS technology (Gielen et al. 2011). Section 6.4.1 first discusses the used topology. The choice of technology determines the nature and magnitude of the different spatial and temporal unreliability

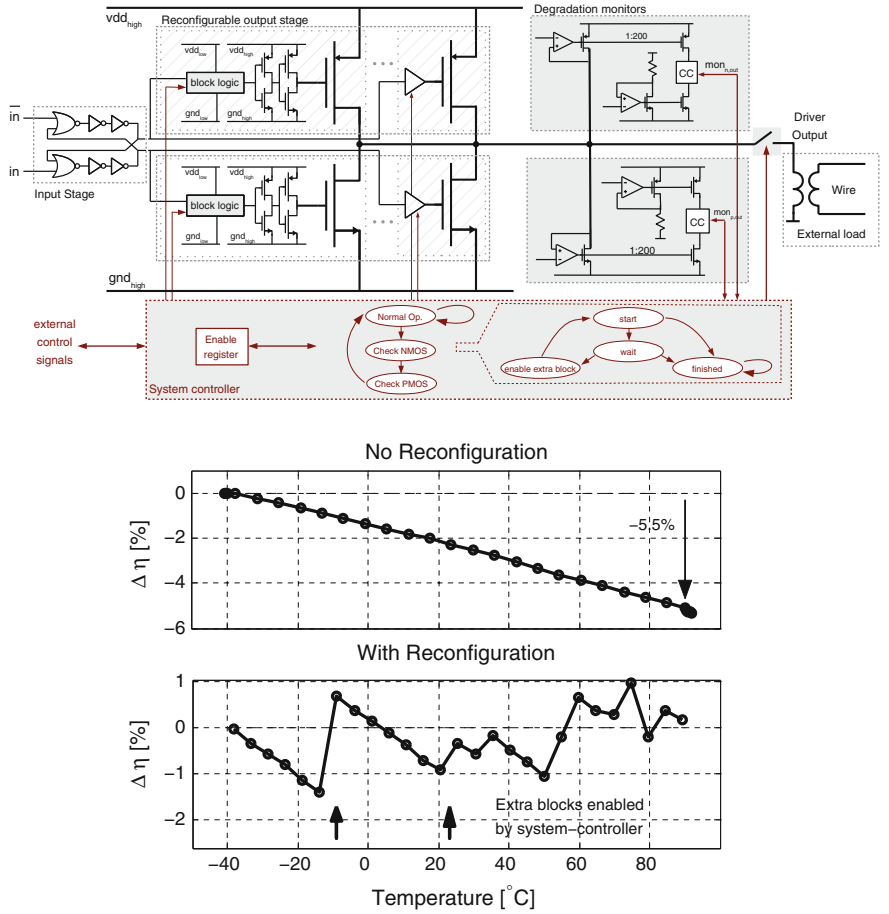


Fig. 6.10 Application of the knobs and monitor principle on a xDSL line driver in a 90 nm CMOS technology (*top*). The power efficiency of the driver is monitored and, as the circuit performance reduces due to aging, extra stages in the output section are switched on, again boosting the efficiency (*bottom*) (De Wit and Gielen 2012)

effects that affect the circuit performance. Then, this technology knowledge is used to design a 6-bit IDAC. Section 6.4.2 first uses using a conventional design flow to guarantee sufficient yield and a low failure rate. This flow uses guardbanding at the device level. However, this typically results in huge circuit overdesign and still does not guarantee reliable circuit operation. In Sect. 6.4.3, the IDAC is redesigned using the reliability models and simulation techniques discussed earlier in this work. These techniques are circuit specific and therefore enable a significant reduction of the design margins, while still guaranteeing or even improving circuit performance. Finally, in Sect. 6.4.4, circuit design techniques are used to further enhance the circuit performance and to reduce the design overhead.

6.4.1 Technology

A predictive 32 nm high-k CMOS technology with 1.1 nm EOT and a nominal $V_{TH} = 0.38$V is used (Arizona state university predictive technology model 2012). The simulation models for each aging mechanism (see Chap. 3) are calibrated with measurements from literature (Cho et al. 2010; Kaczer et al. 2010). Further, the parameter $A_{VTH} = 2.4$mVμm is used to include the impact of process-induced transistor mismatch (Lewyn et al. 2009) (also see Sect. 2.3). The maximum nominal supply voltage $V_{DD,nom}$ to guarantee reliable circuit operation, is typically calculated based on accelerated stress measurements on individual devices. Here, $V_{DD,nom}$ is defined as the stress voltage for which the threshold voltage shift of a single transistor does not exceed a predefined reliability margin of say 50 mV after 5 years. The PBTI, NBTI and SBD aging mechanisms are included in the simulations, since these are considered to be the major failure mechanisms in high-k technologies (also see Chap. 2). The extrapolation from accelerated stress measurements on single devices results in a $V_{DD,nom}$ of 0.91V (see Fig. 6.11). With a $V_{TH} = 0.38$ V, this gives very little headroom to work with, especially when designing an analog circuit where stacked transistors are typically used to achieve a sufficiently large small-signal output resistance. Nevertheless, this is how reliability margins are typically calculated when using a conventional design strategy (Groeseneken et al. 2010). Furthermore, this technique (i.e. reliability assessment based on accelerated stress tests on individual transistors) does not guarantee a reliable circuit. More optimal designs, using higher supply voltages, can be realized if the actual impact of the degradation mechanisms on the circuit is evaluated. The next section first studies the design of an IDAC circuit using a conventional design approach with device-based guardbands. Afterwards, the reliability-aware design flow is applied to the same circuit.

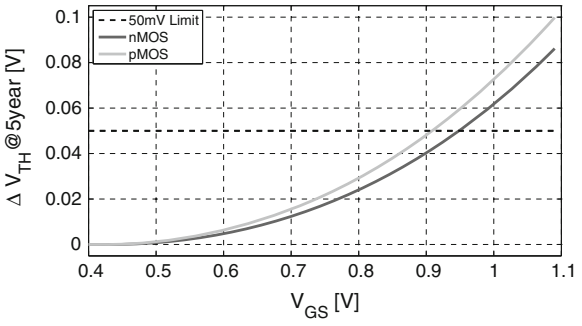


Fig. 6.11 Simulation of the threshold voltage shift after 5 years as a function of the stress voltage for an nMOS and pMOS transistor in a predictive 32 nm CMOS technology with $V_{TH} = 0.38V$. For a reliability margin of $\Delta V_{TH} \leq 50$mV@5years the maximum supply voltage allowed is only 0.91V and is limited by NBTI

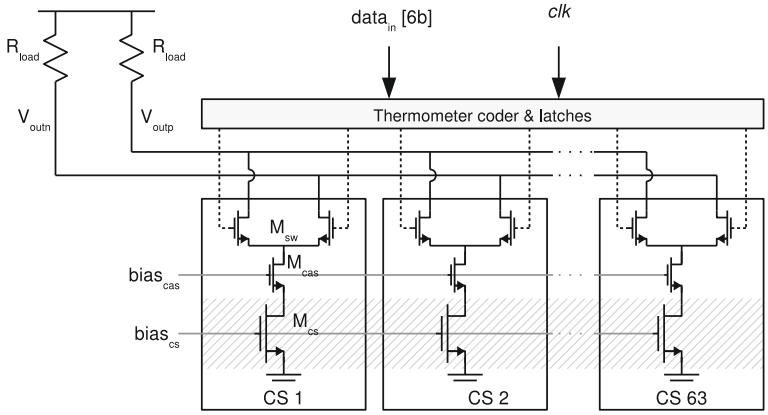


Fig. 6.12 Schematic of a 6-bit current-steering digital-to-analog converter. Reliability simulation is performed on the current-source transistors M_{cs}, shaded in *gray*, which are the accuracy-limiting transistors of the circuit

6.4.2 Conventional Design

The scheme of the demonstrator 6-bit current-steering digital-to-analog converter (IDAC) is depicted in Fig. 6.12. Because of the unary implementation, this IDAC mainly consists of 63 matched unary current-source transistors M_{cs}. Using the switch transistors M_{sw}, the individual currents are routed to one of the two output nodes V_{out+} or V_{out-}, both connected to a fixed load resistor R_{load}. In case of sufficient voltage headroom, cascode transistors M_{cas} are typically added to increase the output impedance. A digital thermometer decoder and clocked latches generate the switch transistor's driving signals, based on the IDAC digital input word $data_{in}$.

From a static performance point of view, the yield of this circuit is limited by the Integral Non Linearity (INL). The INL, defined as the largest difference between the ideal and the actual output value of the DAC across all input codes, should typically be limited to 0.5LSB and is influenced by mismatch between the current-source transistors M_{cs}. Monte-Carlo simulations (Chen and Gielen 2007) are used to determine the maximum allowable current deviation $\sigma(\Delta I_{LSB})/I_{LSB}$, for a certain DAC configuration. Using the Pelgrom mismatch Equations(Pelgrom et al. 1989) and the IDAC specifications, the sizes of the current-source transistors can be calculated. The current source transistors are assumed to be in saturation, therefore a first-order estimation for the required transistor aspect ratio is:

$$\frac{W}{L} = \frac{I_{LSB}}{\beta\,(V_{GS} - V_{TH})^2} \tag{6.2}$$

Process-induced V_{TH} mismatch between two current source transistors results in a mismatch of the output current $\sigma\left(\frac{\delta I_{LSB}}{I_{LSB}}\right)$, with:

$$\sigma\left(\frac{\delta I_{\mathrm{LSB}}}{I_{\mathrm{LSB}}}\right)=\frac{2\sigma\left(\delta V_{\mathrm{TH}}\right)}{V_{\mathrm{GS}}-V_{\mathrm{TH}}} \tag{6.3}$$

Further, the V_{TH} mismatch $\sigma\left(\delta V_{\mathrm{TH}}\right)$ is given by Pelgrom's model (see Eq. 2.1) and can be used to find the minimum required transistor size for a given current source mismatch:

$$WL_{\min}\approx\frac{A_{\mathrm{VTH}}^{2}}{\left[\sigma\left(\frac{\delta I_{\mathrm{LSB}}}{I_{\mathrm{LSB}}}\right)\frac{\left(V_{\mathrm{GS}}-V_{\mathrm{TH}}\right)}{2}\right]^{2}} \tag{6.4}$$

with A_{VTH} the transistor mismatch parameter. According to (6.4), minimal chip area and associated chip cost requires a maximal V_{GS} voltage. On the other hand, preserving the transistor operation in the saturation region limits the usable V_{GS} range: $V_{\mathrm{GS}}\leq V_{\mathrm{DS}}+V_{\mathrm{TH}}\leq V_{\mathrm{DD}}-V_{\mathrm{out,sw,diff}}-V_{\mathrm{DS,cas}}-V_{\mathrm{DS,sw}}-V_{\mathrm{TH}}$. A design value of $V_{\mathrm{GS}}=0.6V_{\mathrm{DD}}$ yields a good compromise.

In the case of high-resolution DACs the area of the unary current source (6.4), and the corresponding current-source matrix area, will dominate the area of the analog part of the digital-to-analog converter. As such, this area is used here to calculate the area-power product. The digital part of the DAC (i.e. the thermometer coder and the latches) is more robust to component variations and aging effects, compared to the analog part, and is therefore not included here.

To guarantee reliable circuit operation, the design flow explained above uses a maximum operating voltage based on reliability measurements on individual transistors (see Sect. 6.4.1). Circuit reliability simulations using the tools explained in Chap. 5, however, indicate that this conventional flow cannot always guarantee reliable circuit operation. For example, when the original circuit is designed for a yield (defined as INL $\leq$ 0.5 LSB and DR $\geq 0.2V_{\mathrm{DD}}$) of 99.7 % (3σ design) and a supply voltage does not exceed the 'safe' $0.91V$ limit, the amount of circuits functioning according to specificiations still reduces to 99.08 % (2.6σ) after 5 years. It can be concluded that reliability assessment based on accelerated stress tests on individual transistors is not a sufficient, nor an appropriate technique for reliable circuit design in advanced high-k technologies. Therefore, in the next sections a design flow using circuit reliability simulation is employed.

6.4.3 Reliability-Aware Design: Fixed Topology

To improve circuit performance, the supply voltage is increased above the 0.91 V limit (see Sect. 6.4.1). Despite the higher V_{DD} and associated increased degradation effects, circuit reliability can still be guaranteed through circuit simulation using the circuit reliability simulation techniques discussed in Chap. 5.

Because of the statistical nature of the degradation effects (see Chap. 2), the spread on the individual current sources M_{cs} can change over time. Starting from Eq. (6.3)

and using the first-order transistor aging model discussed in Sect. 3.6. Equation (6.4) can be extended to also include aging-induced mismatch effects:

$$WL_{\text{min,deg}} \approx \frac{A_{\text{VTH}}^2 + 2A_{\text{BTI}}^2(\Delta V_{\text{TH}})}{\left[\sigma\left(\frac{\delta I_{\text{LSB}}}{I_{\text{LSB}}}\right)\frac{(V_{\text{GS}}-V_{\text{TH}}-\delta V_{\text{TH}})}{2}\right]^2} \tag{6.5}$$

with ΔV_{TH} representing the time-dependent BTI-induced absolute threshold voltage shift and A_{BTI} a technology-dependent parameter. Both ΔV_{TH} and V_{GS} dependent on the applied supply voltage. Therefore, from eq. (6.5), one can assess the impact of a change in V_{DD}. Indeed, if V_{DD} is increased:

- ΔV_{TH} will increase, requiring a *larger* minimum transistor area: $WL_{\text{min,deg}} > WL_{\text{min}}$ since the latter only takes initial process variations into account (see Eq. (6.4)).
- V_{GS} will increase, resulting in a *smaller* minimum area while still realizing the same current accuracy.

In Table 6.2, this trade-off is applied to a single IDAC current source for a 1.0 V and a 1.2 V supply voltage. Operating the current source at 1.2 V results in a five times smaller area; the current mismatch however increases with almost 16 % after 1 year. Figure 6.13 depicts the required area-power product for the entire DAC and for a supply voltage ranging from 0.8 to 1.8 V with a yield target of 99.7 % over circuit lifetime of 5 years (i.e. the black solid line). The square marker in Fig. 6.13 represents the conventional design from the previous Sect. 6.4.2, meeting the yield specifications at design time but not after 5 years of operation. When operating at low supply voltages, the circuit is performance limited by process variations. At higher supply voltages, the circuit performance is limited by PBTI aging effects. Supply voltages higher than 1.4 V strongly increase the probability for hard breakdown events in the

Table 6.2 Reliability analysis of a single DAC current source for a supply voltage of 1.0 V and 1.2 V

V_{DD}	1.0 V	1.2 V	
V_{GST}	0.266 V	0.445 f	**x0.19**
W	0.200 µm	0.060 µm	
L	1.200 µm	0.750 µm	
Area	0.240 µm^2	0.045 µm^2	
I_{DS}	1.999 µA	2.058 µA	
P	2.355 µW	2.688 µW	
ΔV_{TH}@1y	3.7 mV	18.0 mV	**x4.86**
$\sigma(V_{\text{TH}})$@0y	5.18 mV	11.24 mV	
$\sigma(V_{\text{TH}})$@1y	5.19 mV	11.97 mV	
ΔI_{DS}@1y	8.0 nA	0.111 µA	
$\sigma(I_{\text{DS}})/I_{\text{DS}}$@0$y$	2.73 %	2.68 %	
$\sigma(I_{\text{DS}})/I_{\text{DS}}$@1$y$	2.75 %	3.18 %	**x1.16**

Variability parameters are extracted from a simulation on 1000 samples

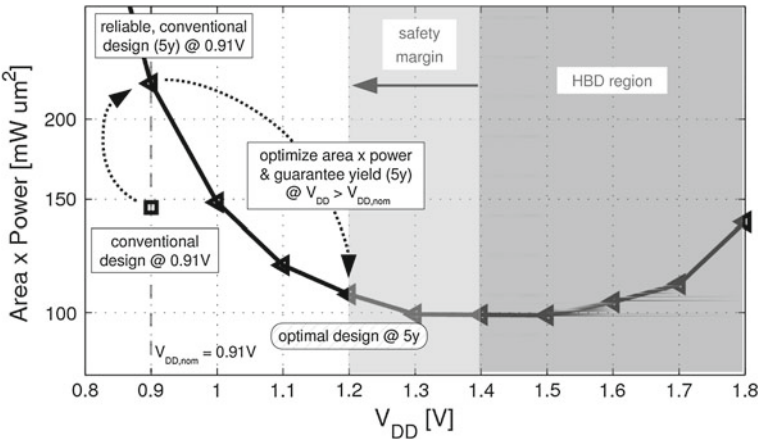


Fig. 6.13 Area-power product of the analog part of the 6-bit IDAC versus the supply voltage. The *black solid line* represents the minimum area-power product for a design with a 99.7 % yield over a lifetime of 5 years. An optimum supply voltage can be found, making optimal use of the 32 nm CMOS technology ($V_{DD} > V_{DD,nom}$) and taking the circuit-specific degradation into account

transistors (also see Sect. 3.4), therefore a 0.2 V backoff is introduced as a safety margin. Eventually, this results in an optimum $V_{DD} = 1.2\,V$ for this circuit, where the area-power product of the analog part of the IDAC is improved by 53 % compared to the design at the nominal supply voltage of 0.91 V, while a 99.7 % yield is still guaranteed over the lifetime of 5 years. As can be imagined, other performance metrics may yield different optimum supply voltages. System-level designers should thus determine the most important performance characteristic(s) and decide on the most appropriate supply voltage accordingly.

6.4.4 Reliability-Aware Design: Digitally-Assisted Analog

The availability of area-efficient, low-power digital circuits in CMOS, allows the implementation of digitally-assisted analog systems. In this way, the effect of performance-limiting analog imperfections can be reduced greatly, leading to designs with a significantly reduced area-power product. Here, the robust IDAC circuit from the previous section is redesigned and digital calibration is used to eliminate various error sources: gradient errors (spatial unreliability effects), gain and offset errors (deterministic unreliability effects), random mismatch induced errors (stochastic unreliability effects). In this section, the example IDAC will be optimized using the switching sequence post adjustment (SSPA) technique (Chen and Gielen 2007). This algorithm optimizes the order in which the individual current sources are switched on with increasing input code, thereby reducing the INL. Originally, the SSPA technique was only applied right after production to correct for process-induced varia-

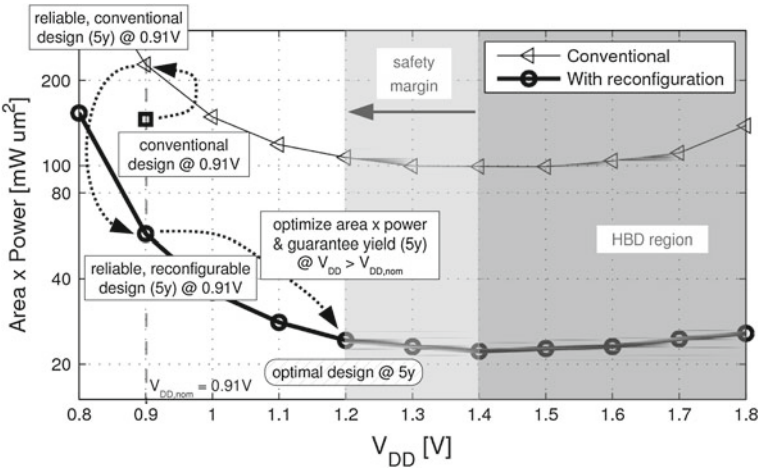


Fig. 6.14 Area-power product of the analog part of a standard and reconfigurable 32 nm CMOS 6-bit IDAC versus the supply voltage. Optimal supply voltages can be found for both implementations. The reconfigurable design (*bold curve*) outperforms the conventional design (*thin curve*) both at the nominal supply voltage as well as at the optimized supply voltage

tions. Here, however, the SSPA algorithm is running at regular time intervals to also correct aging-induced circuit degradation. In Fig. 6.14, the SSPA algorithm is combined with the supply voltage increase technique discussed in the previous section. Application of the SSPA technique strongly improves the performance of the DAC. At the nominal supply voltage (0.91 V), the area-power product decreases by 74 % using the SSPA technique. See (Chen and Gielen 2007) for a detailed explanation and analysis on this area decrease. When also optimizing the operating voltage, based on information from reliability simulations, this results in a supply voltage of 1.2 V and an extra 15 % area-power product decrease. The combination of design techniques (SSPA) and circuit reliability simulations (operating voltage optimization) thus yields an area-power product improvement of 89 % compared to the design at the nominal supply voltage.

6.5 Case Study 2: Digital Circuits

The focus of this work is on analyzing the impact of transistor aging on analog rather than digital circuits. Nevertheless, the latter can be considered as a special type of analog circuits (taking square-shaped waveforms at the input) and can therefore also be analyzed using the models and simulation methods proposed in this work. Given the complexity of a reliability simulation (see Chap. 5), however, only small- to medium-sized digital circuits can be analyzed. Simulation techniques for variability- and reliability-aware analysis of large digital circuits are discussed in (Miranda et al.

2009, Velamala et al. 2011). Below, the analysis of a digital circuit is briefly discussed and demonstrated.

In contradiction to the remainder of this work, this section studies how to determine the lifetime of a *given* circuit rather than analyzing the lifetime at design time. The idea behind this slightly different focus is the increased use of off-the-shelf ICs for the development of cellphones, TVs, cars and even military or space applications. To guarantee reliable operation of the entire system, it is in such cases important to be able to determine the lifetime of each component. This section studies how this can be done, depending on the amount of information available. First, Sect. 6.5.1 overviews the impact of transistor aging on a typical digital circuit. Then, Sect. 6.5.2 discusses how to find a lower bound on the circuit lifetime. Finally, this theory is demonstrated on an example circuit in Sect. 6.5.3.

6.5.1 Digital Circuit Lifetime

As discussed in Sect. 6.2.1, transistor aging can affect a multitude of analog circuit performance parameters. Depending on the application, an analog designer must therefore first identify the cricital circuit performance parameters before he or she can conduct a circuit reliability analysis. In digital circuits, however, typically only two performance parameters are considered: power consumption and speed. The former, could potentially increase over the lifetime of the chip due to SILC effects (Kamohara et al. 1998, Young et al. 2012). Here, the impact of transistor aging on the gate delay is studied.
Figure 6.15 illustrates the time-dependent delay shift for an example logic circuit designed in a (predictive) 32 nm CMOS technology after a stress time of 1 day, 1 month and 1 year. Here, 500 circuit instances were simulated, with each instance subjected to a (different) random input pattern. Similar to an analog circuit (see Sect. 6.2.5), the lifetime of a digital circuit depends a lot on the input activity. Indeed, a different input pattern will activate and age different transistors. Figure 6.15 therefore depicts a distribution of the transistor aging-induced delay shift where the delay shift of one sample depends on the applied input activity. Here, NBTI is the dominant aging effect. Applying a static '001' input therefore results in a worst-case delay shift (i.e. in this case 6 %), while '110' is the best-case input (i.e. no delay shift). Even for this simple example circuit, it is already clear how the input pattern has a big impact on the transistor aging-induced delay shift. Applying a random input pattern to evaluate the lifetime of a circuit is therefore not sufficient to predict aging in a real application. Furthermore, to design a guaranteed reliable system, it is desired to find the minimum circuit lifetime for any applied (realistic) input.

An increase of the path delay, however, does not necessarily result in circuit failure or reduced lifetime (Wang et al. 2007). Most digital circuits are synchronous and therefore, from a circuit or system point of view, the timing conditions provide a relationship between the path delay and the lifetime of the circuit. A circuit is

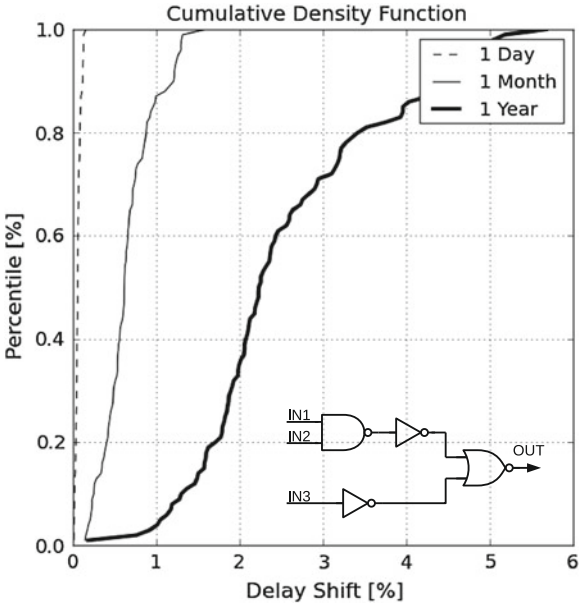


Fig. 6.15 Time-dependent delay shift of an example circuit designed in a (predictive) 32 nm CMOS technology after a stress time of 1 day, 1 month and 1 year. 500 circuit instances are simulated, with each circuit subjected to a random input pattern in time

considered reliable as long as the timing specifications are met:

$$T_{\text{clk}} > t_{\text{p}} + t_{\text{clk,q}} + t_{\text{setup}} - t_{\text{skew}} \tag{6.6}$$

$$t_{\text{skew}} < t_{\text{p}} + t_{\text{clk,q}} - t_{\text{hold}} \tag{6.7}$$

with T_{clk} the clock period, t_{p} the logic path delay and $t_{\text{clk,q}}$ the clock-to-q delay of a flipflop. t_{setup} and t_{hold} are the setup and hold time of a flipflop, respectively. Finally, t_{skew} is the clock skew, which is defined as the clock delay to the destination register minus the clock delay to the source register. Each of the parameters in Eq. (6.6) can be affected by transistor aging. Careful circuit reliability analysis is therefore required. Here, the following definition for lifetime is used: *the lifetime of a digital circuit is the time to the first hold or setup time violation*.

6.5.2 Minimum Circuit Lifetime

Full circuit knowledge (i.e. a circuit netlist) is needed to quantify the actual impact of aging-induced transistor performance shifts on the circuit parameters and to calculate the minimum circuit lifetime. For static CMOS, however, it is also possible to find an

upper bound for the aging-induced *relative* path delay shift, regardless of the circuit architecture and implementation.

A first upper bound for the relative path delay shift can be found when assuming (due to a lack of more circuit knowledge) that every transistor in the circuit ages by a technology-determined maximum amount. For a 32 nm CMOS technology, for example, BTI is the dominant aging mechanism (see Sect. 3.3). BTI results in a shift of the threshold voltage V_{TH} and is primarily a function of the gate-source voltage V_{GS}, the temperature T and the stress time T_{str}:

$$\Delta V_{\mathrm{TH}} = f(V_{\mathrm{GS}}, T_{\mathrm{str}}, T) \tag{6.8}$$

Further, if the gate-source voltage is reduced, part of the BTI damage is recovered (also see Sect. 3.3). Therefore, for a given stress time T_{str}, a given maximum operating temperature T_{max} and a supply voltage V_{DD}, the maximum aging-induced threshold voltage shift $\Delta V_{\mathrm{TH,max}}$ will result from applying a large and fixed stress voltage:

$$\Delta V_{\mathrm{TH,max}} = \Delta V_{\mathrm{TH}}(V_{\mathrm{DD}}, T_{\mathrm{str}}, T_{\mathrm{max}}) \tag{6.9}$$

For example, for a 32 nm CMOS technology with a supply voltage $V_{\mathrm{DD}} = 1.0V$, a lifetime $T_{\mathrm{str}} = 5$ year and a $T_{\mathrm{max}} = 125\,^{\circ}\mathrm{C}$, the maximum transistor threshold voltage shift is 0.15 V (Degraeve et al. 2008; Pae et al. 2008; Lewyn et al. 2009; Cho et al. 2010; Kaczer et al. 2010; Pae et al. 2010). Further, in static CMOS, each transistor is used as a switch. When the circuit is operated, these switches are turned on or off resulting in charges being transported to or from (parasitic) capacitors in the circuit. When switched on, each transistor acts as a current source charging or discharging such a capacitor. The transistor drive current I_{D} is a non-linear function of the transistor gate and drain voltages and is not a constant. Nevertheless, the maximum relative change in transistor drive current can be used as an upper bound for the maximum relative change in circuit delay. When each transistor is assumed to age by a maximum amount:

$$\left.\frac{\Delta t_{\mathrm{p}}}{t_{\mathrm{p}}}\right|_{\mathrm{WC}}^{(T_{str})} \leq \left.\frac{\Delta I_{\mathrm{D}}}{I_{\mathrm{D}}}\right|_{\mathrm{WC}}^{(T_{str})} \tag{6.10}$$

Indeed, the capacitors in the circuit do not change over time and each current source in a path contributes independently to the delay of that path. Eq. (6.10) is only valid for static CMOS circuits, where the ratio between on and off currents is not critical.

A second and more accurate bound on the maximum relative path delay shift can be found for standard cell designs. Most modern LSI circuits are built using standard cells to reduce the design time, to increase yield and to facilitate the design flow. Transistor aging affects the drive current of a standard cell. The aging-induced relative delay change of such a cell $\frac{\Delta t_{\mathrm{g}}}{t_{\mathrm{g}}}$ is therefore a function of the operating voltage V_{DD}, the activity of the input signal α, the slope of the input signal dV_{IN}/dt, the load

capacitance C_{L}, the temperature T and the stress time T_{str}:

$$\frac{\Delta t_{\mathrm{g}}}{t_{\mathrm{g}}} = f(V_{DD}, \alpha, dV_{IN}/dt, C_L, T, T_{\mathrm{str}}) \tag{6.11}$$

Using aging simulations on each gate in the standard cell library and by sweeping over all parameters, one can find the worst-case relative delay change, at time T_{str} and for any gate i in the library:

$$\left.\frac{\Delta t_{\mathrm{g}}}{t_{\mathrm{g}}}\right|_{\mathrm{WC}}^{(T_{\mathrm{str}})} = \max\left(\left.\frac{\Delta t_{\mathrm{g}i}}{t_{\mathrm{g}i}}\right|_{\mathrm{WC}}^{(\mathrm{str})}\right), \forall i \in \{1, \ldots, n\} \tag{6.12}$$

with n the number of gates in the standard cell library. An upper bound for the maximum relative path delay change can then be found by assuming that the worst-case path consists of all worst-case gates:

$$\left.\frac{\Delta t_{\mathrm{p}}}{t_{\mathrm{p}}}\right|_{\mathrm{WC}}^{(T_{\mathrm{str}})} \leq \left.\frac{\Delta t_{\mathrm{g}}}{t_{\mathrm{g}}}\right|_{\mathrm{WC}}^{(T_{\mathrm{str}})} \tag{6.13}$$

6.5.3 Example Circuit

The worst-case lifetime analysis discussed above, is now illustrated on an example circuit, depicted in Fig. 6.16. The circuit is built with a limited standard cell library consisting of a D-flipflop, an AND-gate, an OR-gate and an inverter. Various specifications for each of the standard cells are given in Tables 6.3 and 6.4. Capacitors C_1 to C_4, in Fig. 6.16, represent the interconnect capacitance and the input capacitance to other gates (not shown in the figure). The circuit is operated at a clock frequency of

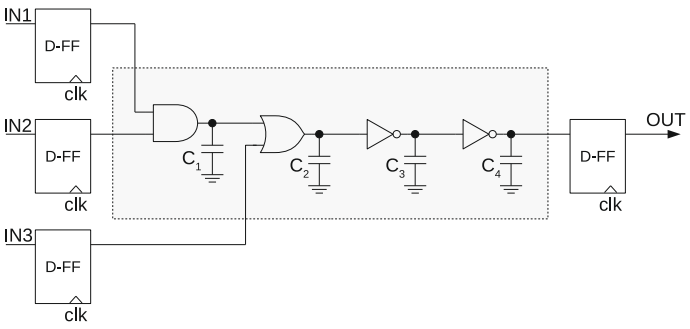


Fig. 6.16 The example datapath circuit to illustrate the aging analysis of a digital circuit

Table 6.3 Gate specifications

Gate	$t_{p,LH}$(ps)	$t_{p,HL}$(ps)	C_{in}(fF)
INV	1.01	2.86	0.32
AND	12.1	9.5	0.35
OR	5.52	27.8	0.35

Table 6.4 Flipflop specifications

	$t_{clk,q}$(ps)	t_{setup}(ps)	t_{hold}(ps)
D-FF	25.0	13.4	0.0

4 GHz. A test pattern to exercise the critical path (i.e. from IN1 to OUT in Fig. 6.16) is also given. At time zero, the critical path delay is 0.185 ns. The circuit has a clock distribution network with a clock-to-register delay $t_{clk} = 7.7$ps.

Section 6.5.2 has discussed two technology-related bounds for the maximum relative path delay shift. As mentioned in Sect. 6.5.1, this delay shift determines the lifetime of the circuit. For a synchronous circuit, that relationship can be found by looking at the setup and hold time conditions (also see Eqs. (6.6) and (6.7)) resulting in a bound on the circuit lifetime. First, one can find a minimum bound on the nominal time-to-failure TTF_n due to an aging-induced setup time violation $TTF_{n,setup}$:

$$\begin{aligned} TTF_{n,setup} &\geq T_{str} \\ &\Updownarrow \\ \frac{T_{clk}}{t_{p,max} + t_{clk,q} + t_{setup} + t_{clk,i}} - 1 &\geq \left.\frac{\Delta t_p}{t_p}\right|_{WC}^{(T_{str})} \end{aligned} \tag{6.14}$$

To find a lower bound on the lifetime without full circuit knowledge (without netlist), one has to assume a worst-case circuit implementation. In that case, the critical path, with delay $t_{p,max}$, ages by a maximum amount. A bound for the maximum relative delay shift was already given in Eqs. (6.10) and (6.13). To find the maximum absolute delay shift of the critical path, $t_{p,max}$ is required. This can either be measured, when a test pattern for the critical path is available, or estimated based on the clock period T_{clk}. Similarly, the clock-to-q delay $t_{clk,q}$ and the setup time t_{setup} of the registers at the input and output of the critical path are also assumed to age by a maximum amount. Both $t_{clk,q}$ and t_{setup} can be obtained from the standard cell library. Finally, the term $t_{clk,i}$ in Eq. (6.14) includes the impact of increased clock skew due to asymmetric stress in the clock distribution network. For a perfectly symmetrical clock distribution network, the skew between the clock signal reaching the input and the output register is zero. However, when circuit techniques such as clock gating are used, different branches of the clock distribution network can age at a different rate. In that case the clock skew can change over time. The worst-case negative clock skew is obtained by assuming that the clock delay to the input register $t_{clk,i}$ changes by a maximum amount while the clock delay to the output register $t_{clk,o}$ remains constant. Circuit imaging techniques such as PICA can be used to extract $t_{clk,i}$ and $t_{clk,o}$ when the

clock distribution network is unknown (Tsang et al. 2000). Solving Eq. (6.14) for T_{str} results in a lower bound for the time to the first setup time violation. Next, similar to $TTF_{n,setup}$, an upper bound for the nominal time-to-failure due to an aging-induced hold time violation $TTF_{n,hold}$ can be found:

$$\begin{aligned} TTF_{n,hold} &\geq T_{str} \\ &\Updownarrow \\ \frac{t_{p,min} + t_{clk,q}}{t_{hold} + t_{clk,o}} - 1 &\geq \left.\frac{\Delta t_p}{t_p}\right|_{WC}^{(T_{str})} \end{aligned} \tag{6.15}$$

Again, solving the equation for T_{str} results in an estimate for the minimum circuit lifetime. Here, the logic path with the smallest possible delay $t_{p,min}$ is observed, because that path will cause the first hold time violation when maximum aging is assumed. $t_{clk,o}$ represents the time for the clock signal to reach the destination register and includes the impact of a maximum positive clock skew.

From Eq. (6.14) and (6.15) one can find an upper limit for the relative path delay shift $\left.\frac{\Delta t_p}{t_p}\right|_{WC}$ to guarantee reliable circuit operation:

$$\left.\frac{\Delta t_p}{t_p}\right|_{WC} \leq \begin{cases} \frac{T_{clk}}{t_{p,max}+t_{clk,q}+t_{setup}+t_{clk}} - 1 = 0.0816 \\ \\ \frac{t_{p,min}+t_{clk,q}}{t_{hold}+t_{clk}} - 1 = 2.246 \end{cases} \tag{6.16}$$

In this example case, the worst-case relative path delay shift should thus not exceed 8.16 %. Based on this limit, various lower bounds on the circuit lifetime can be derived, depending on the available circuit information:

1. First, a lower bound for the circuit lifetime is determined based only on technology information. Figure 6.17 depicts the reliability analysis of a single minimum-sized 32 nm CMOS transistor when it is subjected to a maximum stress voltage (i.e. $V_{GS} = V_{DD}$). The figure shows the relative drain-current shift $\frac{\Delta I_D}{I_D}$, induced by transistor aging, as a function of the stress time and the applied drain-source voltage. Figure 6.17 shows how transistor aging has the biggest impact when the source-drain voltage V_{SD} is large. Using Eqs. (6.10) and (6.16), one can now find a first lower limit on the circuit lifetime: $TTF_{tor} = 3.2$ months (also see Fig. 6.17).
2. A second and more accurate bound can be found when the standard cell library is given. Each standard cell is simulated separately and for each cell the maximum aging-induced delay shift is calculated. Figure 6.18 shows the simulation results for a 32 nm CMOS AND gate. The propagation delay for a low-to-high and high-to-low input pattern is studied, $t_{p,LH}$ and $t_{p,HL}$, respectively. Depending on the applied stress pattern, each of the delays will degrade differently. Various stress patterns were applied and the worst-case $t_{p,LH}$ and $t_{p,HL}$ shift was calculated as a function of the gate fanout and the applied stress time T_{str}. The aging-induced shift of the low-to-high propagation delay is dominant for a small fanout, but for

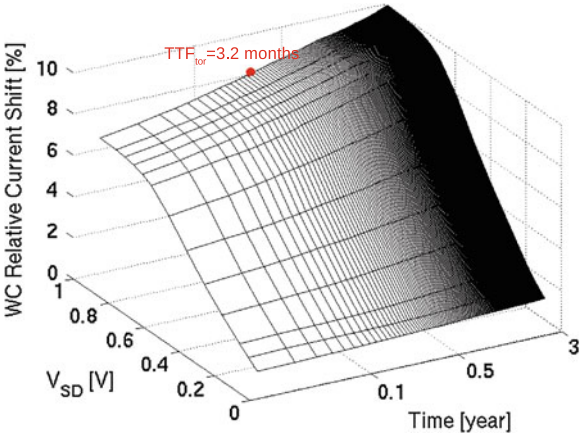


Fig. 6.17 Worst-case drain current shift as a function of the source-drain voltage and the stress time for a minimum-sized 32 nm CMOS transistor

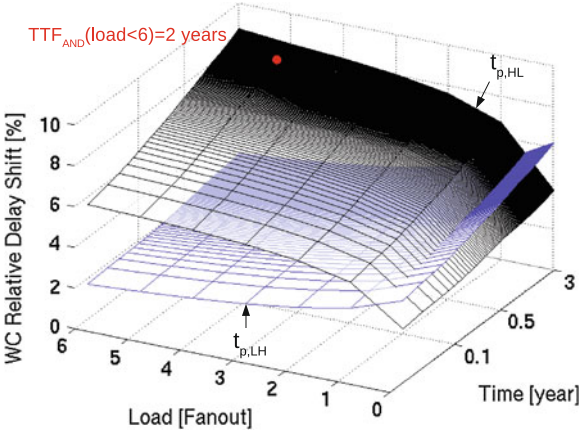


Fig. 6.18 Worst-case relative delay shift of a 32 nm CMOS AND gate as a function of the fanout and the stress time

a large fanout the high-to-low propagation delay shift is larger. Using Fig. 6.18 and Eq. (6.16), an AND-gate time-to-failure TTF_{AND} of 2 years can be found. A similar analysis was done for the other standard cells, resulting in an overall minimum standard cell lifetime of 9.4 months for the inverter gate: $TTF_{std\_cell} = 9.4$ months.

3. Finally, a third lower bound on the circuit lifetime can be calculated when the circuit netlist is given. A reliability analysis of the entire circuit, as a function of the activity of the applied stress patterns, resulted in a worst-case relative delay shift for the entire circuit and as a function of the stress time (see Fig. 6.19). From Fig. 6.19, a lower bound on the circuit lifetime $TTF_{cir} = 2.67$ years can

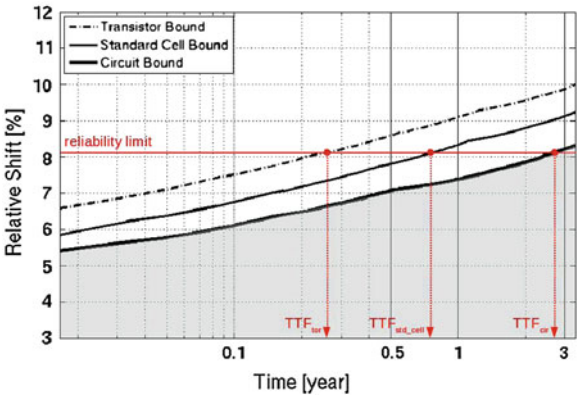


Fig. 6.19 A worst-case aging analysis depending on the available circuit information: based only on transistor aging information, using the standard-cell library and based on the circuit netlist. More information clearly yields a tighter lower bound on the circuit lifetime and a better prediction of the minimum time-to-failure. The *gray* area indicates the relative delay shift of the entire circuit, when subjected to various stress patterns

be derived. Figure 6.19 also shows the worst-case delay shifts resulting from the transistor-only and the standard-cell based approaches. The associated estimates for the time-to-failure are also indicated.

From the analysis of this example circuit, it is clear how technology knowledge can guarantee a rough lower bound for the lifetime of any circuit processed in that technology. But having access to a standard cell library or even to the circuit netlist can drastically improve that bound and make it more realistic. Furthermore, the models and tools proposed in this work can be used to analyze the lifetime of small- to medium-sized digital circuits such as standard cells or small circuit subblocks. A more dedicated approach, as discussed in (Velamala et al. 2011), allows to analyze also larger digital circuits.

6.6 Conclusions

This chapter has discussed the impact of transistor aging on the performance of analog circuits. The lifetime of a circuit has been shown to depend on different aspects: the observed circuit performance parameters, the amount of design guardbanding, the technology, the circuit topology and sizing and the circuit stress conditions. Each of these dependencies has been explained and demonstrated with example circuits. First, an elaborate simulation experiment on various commonly used analog circuit building blocks has revealed the most sensitive circuit performance parameters. Also, each of the test circuits has been designed both in a 65 and a 32 nm CMOS technology. This has shown how more advanced technologies typically bring about larger reliability

problems. Especially BTI is an increasing problem in future technologies. This effect does not only become worse in both nMOS (NBTI) and pMOS (PBTI) transistors but also becomes stochastic, resulting in time-dependent mismatch.

In a second section the knowledge developed in the first section, combined with the models and tools developed in Chap. 5 earlier in this work, have been applied to design a reliable IDAC circuit. It has been shown how a reliability-aware design flow can not only guarantee a more reliable circuit operation but can also help to reduce guardbanding. For the example circuit, the area-power product of the IDAC has been reduced by 89 %, compared to a conventional design, while still having a lifetime of five years. This improvement has been realized with a combination of design techniques and the circuit reliability simulations discussed earlier in this work. Although the focus of this work is on analog circuit reliability simulation, the simulation techniques can also be applied to small- or medium-sized digital circuits. The lifetime of a typical digital standard cell design has been studied and different bounds for the minimum guaranteed lifetime have been derived. Depending on the amount of information available, the bound can be defined more accurately. The theory has also been demonstrated on an example datapath circuit showing a $10\times$ larger minimum lifetime if the full circuit netlist is known, compared to a reliability analysis only based on the technology.

Chapter 7
Conclusions

7.1 General Conclusions

In this work, the impact of transistor aging on analog circuits processed in a conventional nm CMOS process has been investigated. All important transistor aging effects have been studied and compact models for each important effect have been developed. Also, a circuit reliability simulation flow has been proposed. Finally, the flow has been applied to a set of analog circuits and the impact of aging has been studied.

With the scaling to ever-smaller technologies, to reduce cost, area and power consumption and to increase speed, unreliability effects such as process variations and transistor wearout have unfortunately worsened. This requires the use of larger design margins and, combined with an increase in V_{TH} and a reduction of V_{DD}, results in severe reduction of the design space. To counter this problem, a more in-depth and circuit-specific design approach is required. Such a design flow should enable a reduction of the overall circuit guardbands and still guarantee reliable circuit operation by using local overdesign or redundancy.

Three major parts can be distinguished throughout this work:

1. Identification and modeling of transistor unreliability effects.
2. Development of a design for reliability simulation flow.
3. Transistor aging impact analysis in analog ICs.

In the paragraphs below, each of these parts is briefly reviewed and novel techniques and main conclusions are highlighted.

The first part of this work has been covered in Chaps. 2 and 3. First, Chap. 2 has given an overview of the most important unreliability effects in nm CMOS. The effects have been divided in two categories: spatial unreliability effects and temporal unreliability effects. The former are related to process variations and are fixed right after the IC has been processed. The latter are time-dependent and are either transient or permanent. In this work, the impact of permanent transistor aging effects has been studied.

E. Maricau and G. Gielen, *Analog IC Reliability in Nanometer CMOS*,
Analog Circuits and Signal Processing, DOI: 10.1007/978-1-4614-6163-0_7,

Chapter 3 has proposed a set of compact models for each important aging effect: hot carrier injection (HCI), bias temperature instability (BTI) and time dependent dielectric breakdown (TDDB). The models are suited for analog circuit simulation and include all major dependencies and effect properties. Further, an aging-equivalent transistor model to simulate the combined aging effect on the performance of a circuit has been proposed. The results of this part have been published in (Gielen et al. 2008; Maricau et al. 2008, 2011d; Maricau and Gielen 2009b, 2010b).

The second part of the book has focused on developing a reliability simulation flow for analog (and small digital) circuits. This part has been discussed in Chaps. 4 and 5. Chapter 4 has reviewed the most important reliability simulators presented in literature and commercially available. Most of this simulation work has been started in the late 1980s and early 1990s, when electromigration and hot carrier effects reduced circuit reliability. Today, commercial reliability simulation tools such as RelXpert (Cadence) and MOSRA (Synopsis) are still mostly founded on these older approaches. Advanced CMOS technologies, however, suffer from new aging effects such as BTI. Further, stochastic effects such as process variations and BTI in ultra-scaled technologies are becoming more important with each technology node. Prediction of only the nominal circuit lifetime is therefore no longer sufficient and, besides the yield right after production, the circuit failure rate also needs to be analyzed upfront. Chapter 5 has first proposed a deterministic reliability simulator which is capable of analyzing the impact of deterministic aging effects on medium-sized analog or digital circuits. The simulator also includes a sensitivity analysis to detect circuit reliability weak spots. Next, a stochastic reliability simulator has been discussed. This wrapper around the deterministic simulator has added support for process variations and stochastic aging effects. To limit the simulation time, the method uses a response surface method where, based on a limited number of specifically chosen simulations, an analytical circuit model is built. This model is very fast to evaluate and enables fast yield and failure rate predictions. Finally, Chap. 5 has also proposed a hierarchical simulator to simulate the lifetime of large analog circuits. Again a model-based approach has been used to reduce the simulation time. However, here system subblocks are modeled and, besides process-dependent factors, circuit inputs are also part of the model. This severely complicates the sample selection (large input space) and the model itself (non-linear behavior). A new adaptive sample selection algorithm, combined with an advanced symbolic regression model, has been proposed to solve this problem. Each part of the simulator in Chap. 5 has been demonstrated on an example circuit and the simulation results have been verified with hand calculations and analyzed in detail. The simulation methods have been published in (Maricau and Gielen 2009a, 2010a, c, 2011a, b; Gielen et al. 2010; Maricau et al. 2012).

The last part of this work has been covered by Chap. 6 where the models and the simulator developed in the first two parts have been applied on a set of test ICs to evaluate the impact of transistor aging on typical analog circuits. This chapter has first identified and discussed the most important factors that determine the lifetime of an IC: circuit design, technology, stress conditions, the amount of guardbanding and

the performance parameters. Each of these factors has been discussed and illustrated with an example circuit. This improved understanding about the impact of transistor aging on analog ICs can be used by designers as a first step towards designing more reliable circuits. This has been demonstrated on two case studies. First, an IDAC circuit has been analyzed. Here, a combination of a thorough reliability simulation and circuit design techniques has resulted in the design of a reliable circuit with a reduced power-area product. In addition, the method has been applied to the analysis of small digital circuits. Although the proposed models and simulation techniques are intended for analog circuit simulation, they are indeed also applicable to small digital circuits. The results of this part have been published in (Bussche et al. 2011; Gielen et al. 2011; Maricau and Gielen 2011c).

Bibliography

M. Acar, A.-J. Annema, and B. Nauta, Digital detection of oxide breakdown and life-time extension in submicron CMOS technology, in *International Solid-State Circuits Conference*, pp. 530–633, Feb. 2008

K. Agarwal, S. Nassif, Characterizing process variation in nanometer CMOS, in *Design Automation Conference*, pp. 396–399, June 2007

M. Agostinelli, S. Lau, S. Pae, P. Marzolf, H. Muthali, S. Jacobs, PMOS NBTI-induced circuit mismatch in advanced technologies, in *International Reliability Physics Symposium*, pp. 171–175, April 2004

T. Aichinger, M. Nelhiebel, S. Decker, T. Grasser, Energetic distribution of oxide traps created under negative bias temperature stress and their relation to hydrogen. Appl. Phys. Lett. **96**(13), 133511–133511-3 (2010)

M. Alam, A critical examination of the mechanics of dynamic NBTI for PMOSFETs, in *International Electron Devices Meeting*, pp. 14.4.1–14.4.4, Dec. 2003

M. Alam, S. Mahapatra, A comprehensive model of PMOS NBTI degradation. Microelectron. Reliab. **45**(1), 71–81 (2005)

M. Alam, H. Kufluoglu, D. Varghese, S. Mahapatra, A comprehensive model for PMOS NBTI degradation: Recent progress. Microelectron. Reliab. **47**(6), 853–862 (2007)

E. Amat, T. Kauerauf, R. Degraeve, R. Rodriguez, M. Nafria, X. Aymerich, G. Groeseneken, Channel hot-carrier degradation in short channel devices with high-k metal gate stacks, in *Spanish Conference on Electron Devices*, pp. 238–241, Feb. 2009

N. Ampazis, S. Perantonis, Two highly efficient second-order algorithms for training feedforward networks. Trans. Neural Netw. **13**(5), 1064–1074 (2002)

A. Asenov, S. Kaya, A. Brown, Intrinsic parameter fluctuations in decananometer MOSFETs introduced by gate line edge roughness. Trans. Electron Devices **50**(5), 1254–1260 (2003)

A. Asenov, S. Roy, R. Brown, G. Roy, C. Alexander, C. Riddet, C. Millar, B. Cheng, A. Martinez, N. Seoane, D. Reid, M. Bukhori, X. Wang, U. Kovac, Advanced simulation of statistical variability and reliability in nano CMOS transistors, in *International Electron Devices Meeting*, pp. 1–1, Dec. 2008

Arizona state university predictive technology model, http://ptm.asu.edu/, March 2012

S. Aur, D. Hocevar, P. Yang, Circuit hot electron effect simulation. Int. Electron Devices Meet. **33**, 498–501 (1987)

G. Bersuker, D. Heh, C. Young, L. Morassi, A. Padovani, L. Larcher, K. Yew, Y. Ong, D. Ang, K. Pey, W. Taylor, Mechanism of high-k dielectric-induced breakdown of the interfacial SiO2 layer, in *International Reliability Physics Symposium*, pp. 373–378, May 2010

E. Maricau and G. Gielen, *Analog IC Reliability in Nanometer CMOS*,
Analog Circuits and Signal Processing, DOI: 10.1007/978-1-4614-6163-0,

C. Bestory, F. Marc, H. Levi, Statistical analysis during the reliability simulation. Microelectron. Reliab. **47**(9–11), 1353–1357 (2007)
J. Black, Electromigration failure modes in aluminum metallization for semiconductor devices. Proc. IEEE **57**(9), 1587–1594 (1969)
A. Bravaix, C. Guerin, V. Huard, D. Roy, J. Roux, E. Vincent, Hot-carrier acceleration factors for low power management in DC-AC stressed 40 nm NMOS node at high temperature, in *International Reliability Physics Symposium*, pp. 531–548, April 2009
K. Bult, Analog design in deep sub-micron CMOS, in *European Solid-State Circuits Conference*, pp. 126–132, Sept. 2000
S. Bussche, P. De Wit, E. Maricau, G. Gielen, Impact analysis of stochastic transistor aging on current-steering DACs in 32 nm CMOS, in *International Conference on Electronics, Circuits and Systems*, pp. 161–164, Dec. 2011
Cadence, http://www.cadence.com, March 2012
A. Cester, A. Paccagnella, G. Ghidini, S. Deleonibus, G. Guegan, Collapse of MOSFET drain current after soft breakdown. Trans. Device Mater. Reliab. **4**(1), 63–72 (2004)
S. Chakravarthi, A. Krishnan, V. Reddy, C. Machala, S. Krishnan, A comprehensive framework for predictive modeling of negative bias temperature instability, in *International Reliability Physics Symposium*, pp. 273–282, April 2004
S. Chakradhar, A. Raghunathan, Best-effort computing: Re-thinking parallel software and hardware, in *Design Automation Conference*, pp. 865–870, June 2010
I. Chen, S. Holland, C. Hu, Electrical breakdown in thin gate and tunneling oxides. Trans. Electron Devices **32**(2), 413–422 (1985)
T. Chen, G. Gielen, A 14-bit 200-MHz current-steering DAC with switching-sequence post-adjustment calibration. J. Solid-State Circ. **42**(11), 2386–2394 (2007)
M. Cho, M. Aoulaiche, R. Degraeve, B. Kaczer, J. Franco, T. Kauerauf, P. Roussel, L. Ragnarsson, J. Tseng, T. Hoffmann, G. Groeseneken, Positive and negative bias temperature instability on sub-nanometer EOT high-k MOSFETs, in *International Reliability Physics Symposium*, pp. 1095–1098, May 2010
F. Chouard, C. Werner, D. Schmitt-Landsiedel, M. Fulde, A test concept for circuit level aging demonstrated by a differential amplifier, in *International Reliability Physics Symposium*, pp. 826–829, May 2010
F. Chouard, M. Fulde, D. Schmitt-Landsiedel, Reliability assessment of voltage controlled oscillators in 32 nm high-k metal gate technology, in *European Solid State Circuit Conference*, pp. 410–413, Sept. 2010
F. Chouard, S. More, M. Fulde, D. Schmitt-Landsiedel, An aging suppression and calibration approach for differential amplifiers in advanced CMOS technologies, in *European Solid State Circuit Conference*, pp. 251–254, Sept. 2011
P. Clarke, Intel has late metal fix for design error, http://www.eetimes.com/electronics-news/4212726/Intel-has-late-metal-fix-for-design-error
S. Cotter, A screening design for factorial experiments with interactions. Biometrika **66**(2), 317–320 (1979)
F. Crupi, C. Pace, G. Cocorullo, G. Groeseneken, M. Aoulaiche, H.M., Positive bias temperature instability in nMOSFETs with ultra-thin Hf-silicate gate dielectrics. Microelectron. Eng. **80**(6), 130–133 (2005)
A. Davison, D. Hinkley, *Bootstrap Methods and Their Application* (Cambridge University Press, Cambridge, 1997)
K. Deb, *Multi-Objective Optimization Using Evolutionary Algorithms*, vol. 16, (Wiley, Chichester, 2001)
R. Degraeve, M. Aoulaiche, B. Kaczer, P. Roussel, T. Kauerauf, S. Sahhaf, G. Groeseneken, Review of reliability issues in high-k metal gate stacks, in *International Symposium on the Physical and Failure Analysis of Integrated Circuits*, pp. 1–6, July 2008

M. Denais, C. Parthasarathy, G. Ribes, Y. Rey-Tauriac, N. Revil, A. Bravaix, V. Huard, F. Perrier, On-the-fly characterization of NBTI in ultra-thin gate oxide PMOSFETs, in *International Electron Devices Meeting*, pp. 109–112, Dec. 2004

S. Donnay, G. Gielen, *Substrate Noise Coupling in Mixed-Signal ASICs*, (Springer, Berlin 2003)

P. De Wit, G. Gielen, Degradation-resilient design of a self-healing xDSL line driver in 90 nm CMOS. J. Solid-State Circ. **99**, 1–11 (2012)

B. Efron, Bootstrap methods: Another look at the jackknife. Ann. Stat. **7**(1), 1–26 (1979)

Mentor graphics, http://www.mentor.com, March 2012

N. Elias, Acceptance sampling: An efficient, accurate method for estimating and optimizing parametric yield [IC manufacture]. J. Solid-State Circ. **29**(3), 323–327 (1994)

Europractice, http://www.europractice-ic.com/, June 2012

P. Feijoo, T. Kauerauf, M. Toledano-Luque, M. Togo, E. San Andres, G. Groeseneken, Time-dependent dielectric breakdown on subnanometer EOT nMOS FinFETs. Trans. Device Mater. Reliab. **12**(1), 166–170 (2012)

R. Fernández, R. Rodríguez, M. Nafría, X. Aymerich, MOSFET output characteristics after oxide breakdown. Microelectron. Eng. **84**(1), 31–36 (2007)

J. Franco, B. Kaczer, G. Eneman, J. Mitard, A. Stesmans, V. Afanas'ev, T. Kauerauf, P. Roussel, M. Toledano-Luque, M. Cho, R. Degraeve, T. Grasser, L.-A. Ragnarsson, L. Witters, J. Tseng, S. Takeoka, W.-E. Wang, T. Hoffmann, G. Groeseneken, 6A EOT Si0.45Ge0.55 pMOSFET with optimized reliability (VDD=1V): Meeting the NBTI lifetime target at ultra-thin EOT, in *International Electron Devices Meeting*, pp. 4.1.1–4.1.4, Dec. 2010

J. Franco, Multivariate adaptive regression splines. Ann. Stat. 1–67 (1991)

H. Fukutome, Y. Momiyama, T. Kubo, Y. Tagawa, T. Aoyama, H. Arimoto, Direct evaluation of gate line edge roughness impact on extension profiles in sub-50 nm NMOSFETs. Trans. Electron Devices **53**(11), 2755–2763 (2006)

G. Gielen, P. De Wit, E. Maricau, J. Loeckx, J. Martin-Martinez, B. Kaczer, G. Groeseneken, R. Rodriguez, M. Nafria, Emerging yield and reliability challenges in nanometer CMOS technologies, in *Design, Automation and Test in Europe Conference Exhibition*, pp. 1322–1327, March 2008

G. Gielen, E. Maricau, P. De Wit, Design automation towards reliable analog integrated circuits, in *International Conference on Computer-Aided Design*, pp. 248–251, Nov. 2010

G. Gielen, E. Maricau, P. De Wit, Analog circuit reliability in sub-32 nanometer CMOS: Analysis and mitigation, in *Design, Automation Test in Europe Conference Exhibition*, pp. 1–6, March 2011

T. Grasser, R. Entner, O. Triebl, H. Enichlmair, R. Minixhofer, TCAD modeling of negative bias temperature instability, in *International Conference on Simulation of Semiconductor Processes and Devices*, pp. 330–333, Sept. 2006

T. Grasser, W. Gos, V. Sverdlov, B. Kaczer, The universality of NBTI relaxation and its implications for modeling and characterization, in *International Reliability Physics Symposium*, pp. 268–280, April 2007

T. Grasser, W. Goes, B. Kaczer, Dispersive transport and negative bias temperature instability: Boundary conditions, initial conditions, and transport models. Trans. Device Mater. Reliab. **8**(1), 79–97 (2008)

T. Grasser, P.-J. Wagner, P. Hehenberger, W. Goes, B. Kaczer, A rigorous study of measurement techniques for negative bias temperature instability. Trans. Device Mater. Reliab. **8**(3), 526–535 (2008)

T. Grasser, B. Kaczer, Evidence that two tightly coupled mechanisms are responsible for negative bias temperature instability in oxynitride MOSFETs. Trans. Electron Devices **56**(5), 1056–1062 (2009)

T. Grasser, B. Kaczer, W. Goes, H. Reisinger, T. Aichinger, P. Hehenberger, P.-J. Wagner, F. Schanovsky, J. Franco, P. Roussel, M. Nelhiebel, Recent advances in understanding the bias temperature instability, in *International Electron Devices Meeting*, pp. 441–444, Dec. 2010

T. Grasser, B. Kaczer, W. Goes, H. Reisinger, T. Aichinger, P. Hehenberger, P. Wagner, F. Schanovsky, J. Franco, M. Luque, M. Nelhiebel, The paradigm shift in understanding the bias temperature instability: From reaction-diffusion to switching oxide traps. Trans. Electron Devices **58**(11), 3652–3666 (2011)

G. Groeseneken, F. Crupi, A. Shickova, S. Thijs, D. Linten, B. Kaczer, N. Collaert, M. Jurczak, Reliability issues in MuGFET nanodevices, in *International Reliability Physics Symposium*, pp. 52–60, May 2008

G. Groeseneken, R. Degraeve, B. Kaczer, K. Martens, Trends and perspectives for electrical characterization and reliability assessment in advanced CMOS technologies, in *European Solid-State Device Research Conference*, pp. 64–72, Sept. 2010

F. Garcia Sanchez, A. Ortiz-Conde, G. De Mercato, J. Salcedo, J. Liou, Y. Yue, New simple procedure to determine the threshold voltage of MOSFETs. Solid-State Electron. **44**(4), 673–675 (2000)

C. Guerin, V. Huard, A. Bravaix, The energy-driven hot-carrier degradation modes of nMOSFETs. Trans. Device Mater. Reliab. **7**(2), 225–235 (2007)

T. Hastie, R. Tibshirani, J. Friedman, J. Franklin, The Elements of statistical learning: data mining, inference and prediction. Math. Intell. **27**(2), 83–85 (2005)

F. Hong, B. Cheng, S. Roy, D. Cumming, An analytical mismatch model of nano-CMOS device under impact of intrinsic device variability, in *International Symposium on Circuits and Systems*, pp. 2257–2260, May 2011

M. Horstmann, A. Wei, J. Hoentschel, T. Feudel, T. Scheiper, R. Stephan, M. Gerhadt, S. Krugel, M. Raab, Advanced SOI CMOS transistor technologies for high-performance microprocessor applications, in *Custom Integrated Circuits Conference*, pp. 149–152, Sept. 2009

C. Hu, S. Tam, F. Hsu, P. Ko, T. Chan, K. Terrill, Hot-electron induced MOSFET degradation - model, monitor, and improvement. Trans. Electron Devices **32**(2), 375–385 (1985)

C. Hu and Q. Lu, A unified gate oxide reliability model, in *International Reliability Physics Symposium*, pp. 47–51, 1999

V. Huard, C. Parthasarathy, A. Bravaix, T. Hugel, C. Guerin, E. Vincent, Design-in-reliability approach for NBTI and hot-carrier degradations in advanced nodes. Trans. Device Mater. Reliab. **7**(4), 558–570 (2007)

V. Huard, C. Parthasarathy, C. Guerin, T. Valentin, E. Pion, M. Mammasse, N. Planes, L. Camus, NBTI degradation: from transistor to SRAM arrays, in *International Reliability Physics Symposium*, pp. 289–300, May 2008

V. Huard, Two independent components modeling for negative bias temperature instability, in *International Reliability Physics Symposium*, pp. 33–42, May 2010

M. Hussain, R. Barton, S. Joshi, Metamodeling: Radial basis functions versus polynomials. Eur. J. Oper. Res. **138**(1), 142–154 (2002)

IEC 61000 structure, http://www.iec.ch/emc/basic_emc/basic_61000.htm, June 2012

D. Ielmini, M. Manigrasso, F. Gattel, M. Valentini, A new NBTI model based on hole trapping and structural relaxation in MOS dielectrics. Trans. Electron Devices **56**(9), 1943–1952 (2009)

D. Ioannou, S. Mittl, G. La Rosa, Positive bias temperature instability effects in nMOSFETs with HfO2/TiN gate stacks. Trans. Device Mater. Reliab. **9**(2), 128–134 (2009)

Physics of failure in electronics, *in First Annual Symposium on the Physics of Failure in* (Electronics, September, 1962)

International technology roadmap for semiconductors, http://www.itrs.net, 2011

H. Iwai, CMOS technology-year 2010 and beyond. J. Solid-State Circ. **34**(3), 357–366 (1999)

K.O. Jeppson, C.M. Svensson, Negative bias stress of MOS devices at high electric fields and degradation of NMOS devices. J. Appl. Phys. **48**(5), 2004–2014 (1977)

G. Jerke, J. Lienig, Hierarchical current-density verification in arbitrarily shaped metallization patterns of analog circuits. Trans. Comput. Aided Des. Integr. Circ. Syst. **23**(1), 80–90 (2004)

M. Jing, J. Wang, K. Chen, Y. Hao, A study of parametric yield estimation by uniform design sampling. Int. Conf. Solid-State Integr. Circ. Technol. **2**, 1068–1071 (2004)

B. Kaczer, R. Degraeve, M. Rasras, A. De Keersgieter, K. Van de Mieroop, G. Groeseneken, Analysis and modeling of a digital CMOS circuit operation and reliability after gate oxide breakdown: A case study. Microelectron. Reliab. **42**(4–5), 555–564 (2002)

B. Kaczer, A. De Keersgieter, S. Mahmood, R. Degraeve, G. Groeseneken, Impact of gate-oxide breakdown of varying hardness on narrow and wide nFETs, in *International Reliability Physics Symposium*, pp. 79–83, April 2004

B. Kaczer, V. Arkbipov, R. Degraeve, N. Collaert, G. Groeseneken, M. Goodwin, Disorder-controlled-kinetics model for negative bias temperature instability and its experimental verification, in *International Reliability Physics Symposium*, pp. 381–387, 2005

B. Kaczer, T. Grasser, P. Roussel, J. Martin-Martinez, R. O'Connor, B. O'Sullivan, G. Groeseneken, Ubiquitous relaxation in BTI stressing—new evaluation and insights, in *International Reliability Physics Symposium*, pp. 20–27, May 2008

B. Kaczer, P. Roussel, T. Grasser, G. Groeseneken, Statistics of multiple trapped charges in the gate oxide of deeply scaled MOSFET devices—application to NBTI. Electron Device Lett. **31**(5), 411–413 (2010)

B. Kaczer, S. Mahato, V. de Almeida Camargo, M. Toledano-Luque, P. 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 *International Reliability Physics Symposium*, pp. XT.3.1–XT.3.5, April 2011

S. Kamohara, D. Park, C. Hu, Deep-trap SILC (stress induced leakage current) model for nominal and weak oxides, in *International Reliability Physics Symposium*, pp. 57–61, April 1998

E. Karl, P. Singh, D. Blaauw, D. Sylvester, Compact in-situ sensors for monitoring negative-bias-temperature-instability effect and oxide degradation, in *International Solid-State Circuits Conference*, pp. 410–623, Feb. 2008

V. Kaushik, B. O'Sullivan, G. Pourtois, N. Van Hoornick, A. Delabie, S. Van Elshocht, W. Deweerd, T. Schram, L. Pantisano, E. Rohr, L.-A. Ragnarsson, S. De Gendt, M. Heyns, Estimation of fixed charge densities in hafnium-silicate gate dielectrics. Trans. Electron Devices **53**(10), 2627–2633 (2006)

J. Keane, X. Wang, D. Persaud, C. Kim, An all-in-one silicon odometer for separately monitoring HCI, BTI, and TDDB. J. Solid-State Circ. **45**(4), 817–829 (2010)

H.-W. Kim, J.-Y. Lee, J. Shin, S.-G. Woo, H.-K. Cho, J.-T. Moon, Experimental investigation of the impact of LWR on sub-100 nm device performance. Trans. Electron Devices **51**(12), 1984–1988 (2004)

S. Kirkpatrick, C. Gelatt, M. Vecchi, Optimization by simulated annealing. Science **220**(4598), 671 (1983)

A. Krishnan, C. Chancellor, S. Chakravarthi, P. Nicollian, V. Reddy, A. Varghese, R. Khamankar, S. Krishnan, Material dependence of hydrogen diffusion: Implications for NBTI degradation, in *International Electron Devices Meeting*, pp. 691–695, Dec. 2005

H. Kufluoglu, M. Ashraful Alam, A geometrical unification of the theories of NBTI and HCI time-exponents and its implications for ultra-scaled planar and surround-gate MOSFETs, in *International Electron Devices Meeting*, pp. 113–116, Dec. 2004

K. Kuhn, Reducing variation in advanced logic technologies: Approaches to process and design for manufacturability of nanoscale CMOS, in *International Electron Devices Meeting*, pp. 471–474, Dec. 2007

K. Kuhn, C. Kenyon, A. Kornfeld, M. Liu, A. Maheshwari, W. Shih, S.S.G. Taylor, P. Van Der Voorn, K. Zawadzki, Managing process variation in Intel's 45 nm CMOS technology. Intel Technol. J. **12**(2), 93–109 (2008)

K. Lakshmikumar, R. Hadaway, M. Copeland, Characterisation and modeling of mismatch in MOS transistors for precision analog design. J. Solid-State Circ. **21**(6), 1057–1066 (1986)

Y. Leblebici, S.M. Kang, *Hot Carrier Reliability of MOS VLSI Circuits* (Kluwer, Norwell, 1993)

L. Lewyn, T. Ytterdal, C. Wulff, K. Martin, Analog circuit design in nanoscale CMOS technologies. Proc. IEEE **97**(10), 1687–1714 (2009)

X. Li, J. Qin, J. Bernstein, Compact modeling of MOSFET wearout mechanisms for circuit-reliability simulation. Trans. Device Mater. Reliab. **8**(1), 98–121 (2008)
J. Lienig, G. Jerke, Electromigration-aware physical design of integrated circuits, in *International Conference on VLSI Design*, pp. 77–82, Jan. 2005
H. Lipson, J. Bongard, An exploration-estimation algorithm for synthesis and analysis of engineering systems using minimal physical testing, in *Design Automation Conference*, pp. 1087–1093, 2004
M. Lipka, Sony Vaio recall: Poses a burn hazard, http://www.dailyfinance.com/2010/06/30/sony-vaio-recall-poses-a-burn-hazard/, June 2010
J. Loeckx, *Methods for Simulating and Analysing the Effects of EMC on Integrated Circuits*, PhD thesis, (Katholieke Universiteit Leuven, Leuven, Belgium, 2010)
M. Lunenborg, *MOSFET Hot-Carrier Degradation—Failure Mechanisms and Models for Reliability Circuit Simulation*, PhD thesis, (Universiteit Twente, Enschede, Netherlands, 1996)
Y. Luo, J. Orona, D. Nayak, D. Gitlin, Mechanism and modeling of PMOS NBTI degradation with drain bias, in *International Reliability Physics Symposium*, pp. 264–267, April 2007
C. Mack, *Field Guide to Optical Lithography*, (SPIE Press, Bellingham, 2006)
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. Trans. Electron Devices **58**(8), 2347–2353 (2011)
S. Mahapatra, K. Ahmed, D. Varghese, A. Islam, G. Gupta, L. Madhav, D. Saha, M. Alam, On the physical mechanism of NBTI in silicon oxynitride p-MOSFETs: Can differences in insulator processing conditions resolve the interface trap generation versus hole trapping controversy?, in *International Reliability Physics Symposium*, pp. 1–9, April 2007
S. Mahapatra, A. Islam, S. Deora, V. Maheta, K. Joshi, A. Jain, M. Alam, A critical re-evaluation of the usefulness of R-D framework in predicting NBTI stress and recovery, in *International Reliability Physics Symposium*, pp. 614–623, 2011
E. Maricau, P. De Wit, G. Gielen, An analytical model for hot carrier degradation in nanoscale CMOS suitable for the simulation of degradation in analog IC applications. Microelectron. Reliab. **48**(8–9), 1576–1580 (2008)
E. Maricau, G. Gielen, Efficient reliability simulation of analog ICs including variability and time-varying stress, in *Design, Automation Test in Europe Conference Exhibition*, pp. 1238–1241, April 2009
E. Maricau, G. Gielen, A methodology for measuring transistor ageing effects towards accurate reliability simulation, in *International On-Line Testing Symposium*, pp. 21–26, June 2009
E. Maricau, G. Gielen, Variability-aware reliability simulation of mixed-signal ICs with quasi-linear complexity, in *Design, Automation Test in Europe Conference Exhibition*, pp. 1094–1099, March 2010
E. Maricau, G. Gielen, NBTI model for analogue IC reliability simulation. Electron. Lett. **46**(18), 1279–1280 (2010)
E. Maricau, G. Gielen, Efficient variability-aware NBTI and hot carrier circuit reliability analysis. Trans. Comput. Aided Des. Integr. Circ. Syst. **29**(12), 1884–1893 (2010)
E. Maricau, G. Gielen, Computer-aided analog circuit design for reliability in nanometer CMOS. J. Emerg. Sel. Top. Circ. Syst. **1**(1), 50–58 (2011)
E. Maricau, G. Gielen, Stochastic circuit reliability analysis, in *Design, Automation Test in Europe Conference Exhibition*, pp. 1–6, March 2011
E. Maricau, G. Gielen, Transistor aging-induced degradation of analog circuits: Impact analysis and design guidelines, in *European Solid-State Circuit Conference*, pp. 243–246, Sept. 2011
E. Maricau, L. Zhang, J. Franco, P. Roussel, G. Groeseneken, G. Gielen, A compact NBTI model for accurate analog integrated circuit reliability simulation, in *European Solid State Device Research Conference*, 2011

E. Maricau, D. De Jonghe, G. Gielen, Hierarchical analog circuit reliability analysis using multivariate nonlinear regression and active learning sample selection, in *Design, Automation Test in Europe Conference Exhibition*, pp. 745–750, March 2012

T. McConaghy, G. Gielen, Template-free symbolic performance modeling of analog circuits via canonical-form functions and genetic programming. Trans. Circ. Syst. **28**(8), 1162–1175 (2009)

T. McConaghy, FFX: Fast, scalable, deterministic symbolic regression technology, *Genetic Programming Theory and Practice IX*, 2011

J. McPherson, D. Baglee, Acceleration factors for thin gate oxide stressing, in *International Reliability Physics, Symposium*, pp. 1–5, March 1985

M. Miranda, B. Dierickx, P. Zuber, P. Dobrovoln, F. Kutscherauer, P. Roussel, P. Poliakov, Variability aware modeling of SoCs: from device variations to manufactured system yield, in *International Symposium on Quality of Electronic Design*, pp. 547–553, March 2009

J. Martín-Martínez, S. Gerardin, R. Rodríguez, M. Nafría, X. Aymerich, A. Cester, A. Paccagnella, G.G., Lifetime estimation of analog circuits from the electrical characteristics of stressed MOSFETs. Microelectron. Reliab. **47**(9–11), 1349–1352 (2007)

P. Moens, D. Varghese, M. Alam, Towards a universal model for hot carrier degradation in DMOS transistors, in *International Symposium on Power Semiconductor Devices and ICs*, pp. 61–64, June 2010

D. Montgomery, *Design and Analysis of Experiments*, (Wiley, New York, 2008)

B. Murmann, Digitally assisted analog circuits. Micro **26**(2), 38–47 (2006)

L. Nagel, D. Pederson, *SPICE: Simulation Program with Integrated Circuit Emphasis* (University of California, Berkeley, 1973)

A. Neugroschel, G. Bersuker, R. Choi, C. Cochrane, P. Lenahan, D. Heh, C. Young, C. Kang, B. Lee, R. Jammy, An accurate lifetime analysis methodology incorporating governing NBTI mechanisms in *High-k/SiO2 gate stacks, in International Electron Devices Meeting*, pp. 1–4, Dec. 2006

Engineering statistics handbook, http://www.itl.nist.gov/div898/handbook/, April 2012

E. Ogg, HP laptop batteries recalled for overheating, http://news.cnet.com/8301-1001_3-10241137-92.html, May 2009

M. O'Leary, C. Lyden, Parametric yield prediction of complex, mixed-signal ICs. J. Solid-State Circ. **30**(3), 279–285 (1995)

S. Pae, M. Agostinelli, M. Brazier, R. Chau, G. Dewey, T. Ghani, M. Hattendorf, J. Hicks, J. Kavalieros, K. Kuhn, M. Kuhn, J. Maiz, M. Metz, K. Mistry, C. Prasad, S. Ramey, A. Roskowski, J. Sandford, C. Thomas, J. Thomas, C. Wiegand, J. Wiedemer, BTI reliability of 45 nm high-k $+$ metal-gate process technology, in *International Reliability Physics Symposium*, pp. 352–357, May 2008

S. Pae, A. Ashok, J. Choi, T. Ghani, J. He, S. hee Lee, K. Lemay, M. Liu, R. Lu, P. Packan, C. Parker, R. Purser, A. St. Amour, B. Woolery, Reliability characterization of 32 nm high-k and metal gate logic transistor technology, in *International Reliability Physics Symposium*, pp. 287–292, May 2010

C. Parthasarathy, *Etude de la Fiabilite des Technologies CMOS Avancees*, PhD thesis, (Universite d'Aix-Marseille, Marseille, France, 2006)

C. Parthasarathy, M. Denais, V. Huard, G. Ribes, D. Roy, C. Guerin, F. Perrier, E. Vincent, A. Bravaix, Designing in reliability in advanced CMOS technologies. Microelectron. Reliab. **46**(9–11), 1464–1471 (2006)

M. Pelgrom, A. Duinmaijer, A. Welbers, Matching properties of MOS transistors. J. Solid-State Circ. **24**(5), 1433–1439 (1989)

R.G. Phillips, G.P. Anderson, R.A. Erickson, Fundamental failure mechanism studies, in *First Annual Symposium on the Physics of Failure in Electronics*, pp. 73–90, Sept. 1962

Python programming language—official website, http://www.python.org/, April 2012

J. Rabaey, H. DeMan, M. Horowitz, T. Sakurai, J. Sun, D. Dobberpuhl, K. Itoh, P. Magarshack, A. Abidi, H. Eul, Beyond the horizon: the next 10x reduction in power—challenges and solutions, in *International Solid-State Circuits Conference*, p. 31, Feb. 2011

S. Ramey, C. Prasad, M. Agostinelli, S. Pae, S. Walstra, S. Gupta, J. Hicks, Frequency and recovery effects in high-k BTI degradation, in *International Reliability Physics Symposium*, pp. 1023–1027, April 2009

I. Rauch, S.E. G. La Rosa, The energy-driven paradigm of NMOSFET hot-carrier effects. Trans. Device Mater. Reliab. **5**(4), 701–705 (2005)

B. Razavi, *Design of Analog CMOS Integrated Circuits*, (McGraw Hill International edition, USA, 2001)

J. Redoute, *Design of EMI Resisting Analog Integrated Circuits*, PhD thesis, (Katholieke Universiteit Leuven, Leuven, Belgium, 2009)

H. Reisinger, O. Blank, W. Heinrigs, A. Muhlhoff, W. Gustin, C. Schlunder, Analysis of NBTI degradation- and recovery-behavior based on ultra fast VT-measurements, in *International Reliability Physics Symposium*, pp. 448–453, March 2006

H. Reisinger, U. Brunner, W. Heinrigs, W. Gustin, C. Schlunder, A comparison of fast methods for measuring NBTI degradation. Trans. Device Mater. Reliab. **7**(4), 531–539 (2007)

H. Reisinger, T. Grasser, K. Ermisch, H. Nielen, W. Gustin, C. Schlunder, Understanding and modeling AC BTI, in *International Reliability Physics Symposium*, pp. 6A.1.1–6A.1.8, April 2011

Celestry RelXpert, http://www.celestry.com/products_relxpert.shtml, March 2012

Design for reliability: Overview of the process and applicable techniques, http://www.reliasoft.com/newsletter/v8i2/reliability.htm, May 2012

D. Rodopoulos, S. Mahato, V. de Almeida Camargo, B. Kaczer, F. Catthoor, S. Cosemans, G. Groeseneken, A. Papanikolaou, D. Soudris, Time and workload dependent device variability in circuit simulations, in *International Conference on IC Design Technology*, pp. 1–4, May 2011

H.C. Sagong, C.Y. Kang, C.-W. Sohn, M.S. Park, D.-Y. Choi, E.-Y. Jeong, J. Lee, Y.-H. Jeong, New investigation of hot carrier degradation of RF small-signal parameters in high-k/metal gate nMOSFETs, in *International Reliability Physics Symposium*, pp. 5A.4.1–5A.4.5, April 2011

S. Sahhaf, R. Degraeve, P. Roussel, B. Kaczer, T. Kauerauf, G. Groeseneken, A new TDDB reliability prediction methodology accounting for multiple SBD and wear out. Trans. Electron Devices **56**(7), 1424–1432 (2009)

S. Saha, Modeling process variability in scaled CMOS technology. Des. Test Comput. **27**(2), 8–16 (2010)

J. Saleh, K. Marais, Highlights from the early (and pre-) history of reliability engineering. Reliab. Eng. Syst. Saf. **91**(2), 249–256 (2006)

S. Sanchez, P. Sanchez, Very large fractional factorial and central composite designs. Trans. Model. Comput. Simul. **15**(4), 362–377 (2005)

G. Sasse, *Reliability Engineering in RF CMOS*, PhD thesis, (Universiteit Twente, Twente, Netherlands, 2008)

K. Schuegraf, C. Hu, Effects of temperature and defects on breakdown lifetime of thin SiO2 at very low voltages. Trans. Electron Devices **41**(7), 1227–1232 (1994)

D. Schroder, J. Babcock, Negative bias temperature instability: Road to cross in deep submicron silicon semiconductor manufacturing. J. Appl. Phys. **94**(1), 1–18 (2003)

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. Microelectron. Reliab. **45**(1), 39–46 (2005)

Sensitivity analysis, http://en.wikipedia.org/wiki/Sensitivity_analysis, Sept. 2012

B. Settles, Active learning literature survey, Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2009

N. Shanbhag, R. Abdallah, R. Kumar, D. Jones, Stochastic computation, in *Design Automation Conference*, pp. 859–864, June 2010
B. Sheu, W.-J. Hsu, B. Lee, An integrated-circuit reliability simulator-RELY. J. Solid-State Circ. **24**(2), 473–477 (1989)
W. Shockley, Problems related to P-N junctions in silicon. Solid State Electron. **2**(1), 35–67 (1961)
P. Singh, E. Karl, D. Sylvester, D. Blaauw, Dynamic NBTI management using a 45 nm multi-degradation sensor, in *Custom Integrated Circuits Conference*, pp. 1–4, Sept. 2010
P. Solomon, Breakdown in silicon oxide—a review. J. Vac. Sci. Technol. **14**(5), 1122–1130 (1977)
J. Stathis, Physical and predictive models of ultrathin oxide reliability in CMOS devices and circuits. Trans. Device Mater. Reliab. **1**(1), 43–59 (2001)
A. Stefanou, *Analysis and Modelling of the Impact of Substrate Noise on Flash A/D Converters*, PhD thesis, (Katholieke Universiteit Leuven, Leuven, Belgium, 2011)
P. Stolk, F. Widdershoven, D. Klaassen, Modeling statistical dopant fluctuations in MOS transistors. Trans. Electron Devices **45**(9), 1960–1971 (1998)
A. Strojwas, Cost effective scaling to 22 nm and below technology nodes, in *International Symposium on VLSI Technology, Systems and Applications*, pp. 1–2, April 2011
S. Sun, J. Plummer, Electron mobility in inversion and accumulation layers on thermally oxidized silicon surfaces. J. Solid-State Circ. **15**(4), 562–573 (1980)
J. Sune, E. Wu, W. Lai, Statistics of competing post-breakdown failure modes in ultrathin MOS devices. Trans. Electron Devices **53**(2), 224–234 (2006)
J. Suykens, J. Vandewalle, Least squares support vector machine classifiers. Neural Process. Lett. **9**(3), 293–300 (1999)
Synopsys, http://www.synopsys.com, March 2012
E. Takeda, H. Kume, T. Toyabe, S. Asai, Submicrometer MOSFET structure for minimizing hot-carrier generation. Trans. Electron Devices **29**(4), 611–618 (1982)
E. Takeda, N. Suzuki, T. Hagiwara, Device performance degradation to hot carrier injection at energies below the Si-SiO2 energy barrier, in *International Electron Devices Meeting*, pp. 396–399, 1983
K. Takeuchi, T. Fukai, T. Tsunomura, A. Putra, A. Nishida, S. Kamohara, T. Hiramoto, Understanding random threshold voltage fluctuation by comparing multiple fabs and technologies, in *International Electron Devices Meeting*, pp. 467–470, Dec. 2007
S. Tam, P.-K. Ko, C. Hu, Lucky-electron model of channel hot-electron injection in MOSFETs. Trans. Electron Devices **31**(9), 1116–1125 (1984)
M. Toledano-Luque, B. Kaczer, J. Franco, P. Roussel, T. Grasser, T. Hoffmann, G. Groeseneken, From mean values to distributions of BTI lifetime of deeply scaled FETs through atomistic understanding of the degradation, in *Symposium on VLSI Technology*, pp. 152–153, June 2011
J.C. Tsang, J.A. Kash, D.P. Vallett, Picosecond imaging circuit analysis. IBM J. Res. Dev. **44**(4), 583–603 (2000)
R. Tu, E. Rosenbaum, W. Chan, C. Li, E. Minami, K. Quader, P. Ko, C. Hu, Berkeley reliability tools—BERT. Trans. Comput. Aided Des. Integr. Circ. Syst. **12**(10), 1524–1534 (1993)
K. Tu, Recent advances on electromigration in very-large-scale-integration of interconnects. J. Appl. Phys. **94**(9), 5451–5473 (2003)
B. Tudor, J. Wang, Z. Chen, R. Tan, W. Liu, F. Lee, An accurate and scalable MOSFET aging model for circuit simulation, in *International Symposium on Quality, Electronic Design*, pp. 1–4, March 2011
A. Vassighi, O. Semenov, M. Sachdev, A. Keshavarzi, C. Hawkins, CMOS IC technology scaling and its impact on burn-in. Trans. Device Mater. Reliab. **4**(2), 208–221 (2004)
R. Van De Plasche, *Integrated Analog-to-Digital and Digital-to-Analog Converters* (Kluwer, Boston, 1994)
J. Velamala, V. Ravi, Y. Cao, Failure diagnosis of asymmetric aging under NBTI, in *International Conference on, Computer-Aided Design*, pp. 428–433, Nov. 2011

W. Wang, V. Reddy, A. Krishnan, R. Vattikonda, S. Krishnan, Y. Cao, Compact modeling and simulation of circuit reliability for 65 nm CMOS technology. Trans. Device Mater. Reliab. **7**(4), 509–517 (2007)

W. Wang, S. Yang, S. Bhardwaj, S. Vrudhula, F. Liu, Y. Cao, The impact of NBTI effect on combinational circuit: Modeling, simulation, and analysis. Trans. Very Large Scale Integr. Syst. **18**(2), 173–183 (2010)

X. Wang, A. Brown, B. Cheng, A. Asenov, Statistical variability and reliability in nanoscale FinFETs, in *International Electron Devices Meeting*, pp. 5.4.1–5.4.4, Dec. 2011

Wilcoxon signed rang test, http://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test, March 2012

D.R. Wolters, J.J. van der Schoot, Kinetics of charge trapping in dielectrics. J. Appl. Phys. **58**(2), 831–837 (1985)

H. Wong, M. Poon, Approximation of the length of velocity saturation region in MOSFETs. Trans. Electron Devices **44**(11), 2033–2036 (1997)

E. Wu, J. Sune, W. Lai, E. Nowak, J. McKenna, A. Vayshenker, D. Harmon, Interplay of voltage and temperature acceleration of oxide breakdown for ultra-thin oxides. Microelectron. Eng. **59**(1–4), 25–31 (2001)

E. Wu, J. Sune, W. Lai, E. Nowak, J. McKenna, A. Vayshenker, D. Harmon, Interplay of voltage and temperature acceleration of oxide breakdown for ultra-thin gate oxides. Solid-State Electron. **46**(11), 1787–1798 (2002)

E. Wu, J. Su, Power-law voltage acceleration: A key element for ultra-thin gate oxide reliability. Microelectron. Reliab. **45**(12), 1809–1834 (2005)

E. Wu, R. Vollertsen, J. Sune, G. La Rosa, *Reliability Wearout Mechanisms in Advanced CMOS Technologies*, vol. 12, (IEEE Press, Wiley, 2009)

Xbox_360_technical_problems, http://en.wikipedia.org/wiki/Xbox_360_technical_problems, May 2012

X. Xuan, A. Chatterjee, A. Singh, ARET for system-level ic reliability simulation, in *International Reliability Physics Symposium*, pp. 572–573, 2003

K. Yang, B. El-Haik, *Design for Six Sigma* (McGraw-Hill, New York, 2003)

B. Yan, Q. Fan, J. Bernstein, J. Qin, J. Dai, Reliability simulation and circuit-failure analysis in analog and mixed-signal applications. Trans. Device Mater. Reliab. **9**(3), 339–347 (2009)

S. Yazdani, Electrostatic discharge (ESD) explained, http://www.electronicspub.com/article/26/7/ElectroStatic-Discharge-(ESD)-Exp

C. Young, G. Bersuker, M. Jo, K. Matthews, J. Huang, S. Deora, K. Ang, T. Ngai, C. Hobbs, P. Kirsch, A. Padovani, L. Larcher, New insights into SILC-based life time extraction, in *International Reliability Physics Symposium*, pp. 5D.3.1–5D.3.5, April 2012

S. Zafar, Statistical mechanics based model for negative bias temperature instability induced degradation. J. Appl. Phys. **97**(10), 103709–103709-9 (2005)

L. Zhang, J. Gove, L. Heath, Spatial residual analysis of six modeling techniques. Ecological Modelling **186**(2), 154–177 (2005)

W. Zhao, Y. Cao, F. Liu, K. Agarwal, D. Acharyya, S. Nassif, K. Nowka, Rigorous extraction of process variations for 65 nm CMOS design, in *European Solid State Circuits Conference*, pp. 89–92, Sept. 2007

L. Zhang, J. Zhou, J. Im, P. Ho, O. Aubel, C. Hennesthal, E. Zschech, Effects of cap layer and grain structure on electromigration reliability of Cu/low-k interconnects for 45 nm technology node, in *International Reliability Physics Symposium*, pp. 581–585, May 2010

Index

E. Maricau and G. Gielen, *Analog IC Reliability in Nanometer CMOS*, Analog Circuits and Signal Processing, DOI: 10.1007/978-1-4614-6163-0,