Reliability Aware Computing Platforms Design and Lifetime Management

Cucu Laurenciu, Nicoleta

DOI
10.4233/uuid:9c46dcd4-e68e-42de-a7ab-676a3e631f0d

Publication date
2017

Document Version
Final published version

Citation (APA)
Cucu Laurenciu, N. (2017). *Reliability Aware Computing Platforms Design and Lifetime Management*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:9c46dcd4-e68e-42de-a7ab-676a3e631f0d

Reliability Aware Computing Platforms
Design and Lifetime Management

Cover illustration: Constantin Brancusi's monumental Endless Column in Targu-Jiu, Romania. The column of metal-coated cast-iron modules on a steel spine with a height of 29.33 meters, is part of a sculptural ensemble originally installed in 1937-1938, which includes the travertine Table of Silence and Gate of the Kiss, as well.

Reliability Aware Computing Platforms Design and Lifetime Management

---

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. K.C.A.M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op

donderdag 26 januari 2017, om 10:00 uur

door

Nicoleta CUCU LAURENCIU

Master of Science in Computer Engineering
Technische Universiteit Delft, Nederland
geboren te Brasov, Romania

Dit proefschrift is goedgekeurd door de promotor:
Prof. dr. ir. K. L. M. Bertels

Copromotor:
Dr. S. D. Cotofana

Samenstelling promotiecommissie:

Rector Magnificus, voorzitter	Technische Universiteit Delft
Prof. dr. ir. K. L. M. Bertels	Technische Universiteit Delft, promotor
Dr. S. D. Cotofana	Technische Universiteit Delft, copromotor

Onafhankelijke leden:

Prof. dr. ir. G. Leus	Technische Universiteit Delft, Nederland
Prof. dr. A. Rubio	Universitat Politècnica de Catalunya, Spain
Dr. C. Anghel	Institut Supérieur d'Électronique de Paris, France
Prof. dr. ir. K. Goossens	Technische Universiteit Eindhoven, Nederland
Prof. dr. G. Stefan	"Politehnica" University of Bucharest, Romania
Prof. dr. ir. P. Van Mieghem	Technische Universiteit Delft, reserve lid

CIP-DATA KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Nicoleta Cucu Laurenciu

Reliability Aware Computing Platforms Design and Lifetime Management
Delft: TU Delft, Faculty of Elektrotechniek, Wiskunde en Informatica - III
Thesis Technische Universiteit Delft.

Met samenvatting in het Nederlands.

ISBN 978-94-6186-780-3

Keywords: Reliability, Reliability Aware Computation, Dynamic Lifetime Reliability Management, Reliability Assessment

Printed in Nederland

I dedicate the thesis to my mentor

Abstract

Aggressive CMOS technology feature size down-scaling into the deca nanometer regime, while benefiting performance and yield, determined device characteristics variability increase w.r.t. their nominal values, which can lead to large spreads in delay, power, and robustness, and make devices more prone to aging and noise induced failures during in-field usage. Because of transistor's gate dielectric increasing power density and electric field the nanoscale Integrated Circuits (ICs) failure mechanisms accelerating factors have become more severe than ever, which can cause higher failure rate during ICs useful life and early aging onset. As a result, meeting the reliability targets with viable costs in this landscape becomes a significant challenge, requiring to be addressed in an unitary manner from design time to run time. To this end, we propose a holistic reliability aware design and lifetime management framework concerned (i) at design time, with providing a reliability enhanced adaptive architecture fabric, and (ii) at run time, with observing and dynamically managing fabric's wear-out profile such that user defined Quality-of-Service requirements are fulfilled, and with maintaining a full-life reliability log to be utilized as auxiliary information during the next IC generation design. Specifically, we first introduce design time transistor and circuit level aging models, which provide the foundation for a 4-dimensional Design Space Exploration (DSE) meant to identify a reliability optimized circuit realization compliant with area, power, and delay constraints. Subsequently, to enable the creation of a low cost but yet accurate fabric observation infrastructure, we propose a methodology to minimize the number of aging sensors to be deployed in a circuit and identify their location, and introduce a sensor design able to directly capture circuit level amalgamated effects of concomitant degradation mechanisms. Furthermore, to make the information collected from sensors meaningful to the run-time management framework we introduce a circuit level model that can estimate the overall circuit aging and predict its End-of-Life based on imprecise sensors measurements, while taking into account the degradation nonlinearities. Finally, to provide more DSE reliability enhancement options we focus on the realization of reliable data transport and processing with unreliable components, and propose: (i) a codec for reliable energy efficient medium/long range data transport, and (ii) a methodology to obtain Error Correction Codes protected data processing units with an output error rate smaller than the fabrication technology gate error rate.

Acknowledgments

The past 5 years work whose end result is constituted by this thesis, was enabled by the contribution of several excellent people, to whom I would like to pay my respects chronologically.

I would like to thank my former professors from Transilvania University of Brasov, Romania, for laying the Computer Engineering (CE) knowledge foundations. I would like to thank especially Prof. dr. Gheorghe Toacse, founder of the CE department at Transilvania University of Brasov, one of the Romanian university education titans, who introduced me to the rigors of computer engineering and laid the foundations of structured thinking in this domain. His influence on my knowledge and professional formation was significant. He also facilitated me to begin my studies in Delft, as a terminal year bachelor exchange student, and I am forever grateful.

I thank my former professors from TU Delft, The Netherlands, for consolidating and enriching my knowledge during the M.Sc. studies in Computer Engineering.

Foremost I would like to express my deepest, utmost and sincere gratitude to my Ph.D. supervisor, Dr. Sorin D. Cotofana, for shaping and cultivating the quality of my mind and character, the freedom of thought and for awakening an ardent spirit of inquiry, while instilling at the same time enduring life values. He permanently guided me while allowing me complete freedom of exploration, he taught me the modus operandi of critical thinking and not what to think, and facilitated me to experience knowledge for the sake of knowledge. His passion and inexorable dedication for research and academia earned my profound esteem. Prof. Cotofana incessantly motivated me to keep progressing, and encouraged assiduous and disciplined pursuit of knowledge. His moral stamina and energy were always a source of inspiration. In any situation he was always objective, and with an irreproachable moral stance. My gratitude toward Prof. Cotofana is limitless. It was truly a privilege, a pleasure and an honor to work and interact with him.

I am thankful to Prof. dr. ir. Koen Bertels, the head of the Quantum Computing Department, who brought a welcomed influence, by supporting us in our academic development, encouraging a unified collaboration working atmosphere, and creating a pleasant intellectual environment in the department. He broadened our horizon in other scientific areas via the various topics colloquia regularly held by top field professors.

I would like to thank the CE secretary Ms. Lidwina Tromp for guiding me through the

administrative and financial cobweb, multum in parvo for her promptitude, availability and professionalism. I have enjoyed the pleasant conversations with her and greatly benefited from her advice on various matters as well.

I would like to extend my thanks to Dr. ir. Arjan van Genderen for his support, enabling me to be independent to a certain extent, which positively influenced my EDA tools knowledge and simulations.

I am grateful to the technical support team, foremost Erik de Vries, the CE system administrator, for ensuring a fluent and smooth operation of the simulation infrastructure on the HPC clusters, a sine qua non component of research infrastructure.

Within the framework of i-RISC international project, I had the opportunity to collaborate with Dr. Valentin Savin, from CEA-LETI, France. I was impressed by his work conduit and his clear, structured mathematical thinking, and I thank him for the fruitful discussions concerning LDPC codes and their utilization in reliable functional units design.

I would also like to thank the members of the committee for accepting to take part in the defense committee and for reviewing the thesis.

I would like to thank the Ph.D. colleagues, especially the Chinese ones, for the fruitful exchange of ideas.

I also thank my parents, of whom material and moral support I enjoyed throughout my studies.

Nicoleta Cucu Laurenciu Delft, The Netherlands, January 2017

Table of Contents

List of Tables

List of Figures

List of Acronyms

AIC	Akaike Information Criterium
BICS	Built-In Current Sensor
CAD	Circuit Assisted Design
CLM	Channel Length Modulation
CMOS	Complementary Metal-Oxide-Semiconductor
CPD	Current-mode Peak Detector
C2T	Current-to-Time Converter
CUO	Circuit Under Observation
DPU	Data Processing Unit
DRM	Dynamic Reliability Management
DRM	Dynamic Reliability Management
ECC	Error Correcting Code
ECL	Emmiter-Coupled Logic
EM	Electro-Migration
EOL	End-Of-Life
FEOL	Front-End-Of-Line
FET	Field Effect Transistor
GDP	Gross Domestic Product
HCI	Hot Carrier Injection
HDL	Hardware Description Language
IC	Integrated Circuit
KL	Karhunen Loeve
LP	Low Power
LSB	Least Significant Bit
MD	Mobility Degradation
MOS	Metal-Oxide-Semiconductor
MOSFET	Metal-Oxide-Semiconductor Field Effect Transistor
MR	Modular Redundancy
MSB	Most Significant Bit
NBTI	Negative Bias Temperature Instability
NoC	Network on Chip
NMOS	N-type Metal-Oxide-Semiconductor
NWE	Narrow Width Effects
PBTI	Positive Bias Temperature Instability
PMOS	P-type Metal-Oxide-Semiconductor
PDF	Probability Density Function
PI	Primary Input
PO	Primary Output
PWM	Pulse Width Modulated
RDS	Source-Drain parasitic Resistance

RM	Reed-Müller
SCE	Short Channel Effects
SECDED	Single Error Correction Double Error Detection
SoC	System on Chip
STA	Statistical Timing Analysis
SSTA	Statistical Static Timing Analysis
TC	Thermal Cycling
TDDB	Time Dependent Dielectric Breakdown
TTL	Transistor-Transistor Logic
VS	Velocity Saturation
WER	Word Error Rate

1 Introduction

Integrated Circuits (ICs) brought about a computerization and electronics revolution, which radically changed the preponderant manner of human living and interaction. The silicon ICs age has its roots marked by three salient milestones: (i) the Field Effect Transistor (FET) concept introduction in 1926 by Julius Edgar Lilienfeld [1], and in 1934 by Oskar Heil [2], (ii) the invention of the bipolar transistor at Bell Telephone Laboratories by John Bardeen, Walter Brattain, and William Shockley in 1947 [3], and (iii) the introduction in 1959 of the planar manufacturing process by Jean Hoerni (Fairchild Semiconductor) [4], and of the monolithic approach by Robert Noyce (Fairchild Semiconductor) [5], which revolutionized the semiconductor manufacturing and technology.

One year later, the ICs era started, as the first working monolithic ICs were produced, credited to Jack Kilby (Texas Instruments) and Robert Noyce (Fairchild Semiconductor) [6]. Early 1960's witnessed the low density fabrication process with only a few transistors per chip, known as Small Scale Integration (SSI), which paved the way to the Medium Scale Integration (MSI) at the end of the decade, with a chip transistor count of 100 or more. Military and aerospace computer systems were among the first to benefit from the ICs applications. As more competitors entered the market share race, multiple bipolar IC standard logic families were pioneered, e.g., Emitter-Coupled Logic (ECL) [7], Transistor-Transistor Logic (TTL) [8]. ECL which originated in late 1950's at IBM, was the fastest logic family available (with picoseconds logic gate delays), but suffered from low noise margin and high power dissipation, i.e., $3\times$ to $10\times$ higher than that of the TTL logic family. Patented in 1961 by James Buie (Pacific Semiconductor), TTL had the advantage of allowing for higher integration density and of being less power hungry. By the late 1960's, the TTL logic family and its successors established as the logic market share leaders for nearly 2 decades, being the impetus for the semiconductor industries evolution. In the same time, the Metal-Oxide Semiconductor Field Effect Transistor (MOSFET) entered production and in 1964, the first commercial Metal-Oxide Semiconductor (MOS) ICs appeared [6]. As a matter of fact, despite the FET principle being invented in 1926 and 1934, due to material related problems, early attempts to fabricate functioning devices were hindered for nearly 4 decades. Early commercial MOS processes employed pMOS transistors

only, which had the drawbacks of low performance and reliability, while in the 1970's the nMOS-only logic took over due to its higher speed. Intel inaugurated in 1972 the nMOS technology with the first mass-producible microprocessor (the 4-bit 4004, containing 2300 transistors on an area of only 3 by 4 millimetres, which consumed tenths of a Watt, and had a price of $\approx 200US\$$) [9]. Programmable intelligence was now attainable with an affordable cost, and Intel 4004 conquered the market within months after its introduction.

In the 1980's, as Very-Large-Scale Integration (VLSI) process enabled ICs to hold hundreds of thousands of transistors, power consumption aroused as a dominant issue. In this context, the MOS family won the hegemony over previous bipolar ECL and TTL logic families (though still in use today in special very high performance demanding situations, e.g., BiCMOS that combines bipolar and MOS transistors in high-speed memories, ECL successors in high frequency digital communications), mainly for the same reasons vacuum tubes were deemed inappropriate and were replaced by bipolar transistors, i.e., the high power dissipation, which imposed a hard limit on the integration density envelope. From the MOS family, the CMOS technology was widely adopted and preferred over nMOS-only and pMOS-only, mainly due to the manufacturing process simplicity and its "zero" static power consumption characteristic, becoming the dominant technology for nearly all logic applications.

Since 1970's MOS ICs integration density and performance underwent a staggering revolution as (i) transistors count per die grew exponentially, doubling roughly each and every 18 months, reaching billions of transistors in today's multi-core processors, and (ii) ICs performance had a similar evolution trend, e.g., the clock frequency of Intel's microprocessors doubled approximately every 3 years. This exponential growth was very much in line with the insightful prediction Gordon Moore (Fairchild Semiconductor) formulated in 1965, which became known as "Moore's law" and prevailed for the past 5 decades [10]. "Moore's law" was adopted as target guideline by the semiconductor companies to maintain their market position w.r.t. competitors. In 1974, Robert Dennard (IBM) formulated a corollary to the "Moore's law" [11], which states that as MOSFET transistors become smaller, they consume less power and are cheaper to manufacture. As the number of transistors per chip increased from hundreds/cm^2 to billions/cm^2, and the cost per transistor decreased with roughly 29% per year to reach less than 0.05US\$, ICs spread to virtually all economy sectors, from defence industry, aerospace, telecommunications, power, to automotive, health, and food industries, and the semiconductor industry became a strategic driver of the global economy growth, accounting for a large percentage of the total productivity gains (e.g., close to 10% of the world Gross Domestic Product (GDP) [12]).

However, as the CMOS manufacturing process feature size continues to shrink, despite the individual transistors consuming very little power per switching, billions of them switching at very high speed and leaking as well, brought back once more a power wall to further miniaturization. Moreover, with feature size reduction, the manufacturing process becomes less precise as increasingly smaller geometries should be handled, more defects are induced, and transistors become more susceptible to various in-field failures. As a result, transistors no longer operate as expected, their

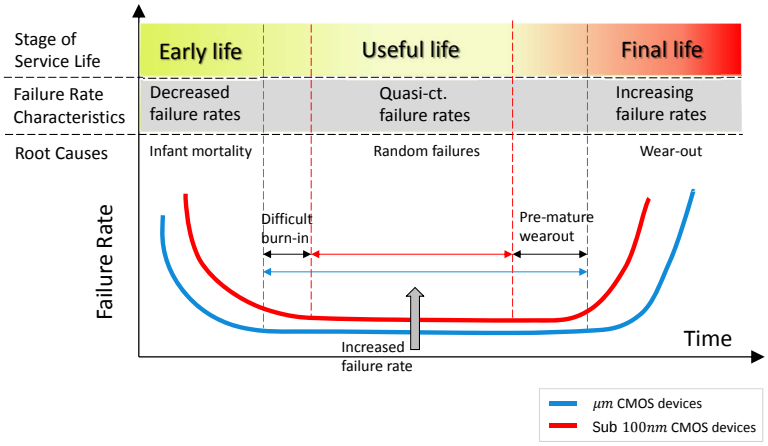


Figure 1.1: Lifetime Failure Distribution for Advanced Technology Nodes.

performance is poorer with respect to the nominal one, or they even fail altogether pre-maturely, i.e., within their intended useful life. The decreasing device reliability poses an increasing concern for current technology nodes, and can potentially become a showstopper to further CMOS scaling.

In view of the reliability wall imminence at the technology horizon, this thesis addressees nano device, circuit, and architecture reliability issues erected by the aggressive technology scaling, as subsequently detailed.

1.1 ICs Reliability Trends

According to IEEE, reliability is defined as the ability of a product, in particular of an IC, to perform as intended (without failures and within intended performance envelopes) for a specified period of time (i.e., during the product lifetime). A product experiences a failure, if the service/performance delivered is not consistent with the expected datasheet service/performance.

During its lifetime, an IC can experience various failures during different life stages. Historically, the IC failure rate during its lifetime, is empirically modeled by the ”bathtub” curve [13]. Figure 1.1 illustrates typical ”bathtub” curves for a large population of identical and independent ICs, implemented in micro and nano technology generations, and subjected to identical stress conditions. The ”bathtub” curve consists of 3 characteristic regions, as follows. During the first stage, i.e., at the beginning of ICs life, the failure rate, which is typically associated with manufacturing defects (e.g., oxide defects) is rapidly decreasing. The ICs suffering from congenital defects that escaped time-zero and yield tests are responsible for the high early life failure rate.

Stress screening (burn-in) proved to be effective in detecting and removing those infant defective ICs such that only the healthy ones, i.e., that passed the burn-in tests, are shipped to customers and enter into the second "bathtub" curve stage, i.e., the useful life, which is characterized by a small and quasi-constant failure rate. During this period, failures are random, and intrinsic to the environmental conditions (e.g., radiation, noise), manifesting themselves mostly as transient (soft) errors. Typically, reducing the in-field stress conditions to which ICs are exposed, and/or increasing the design robustness (e.g., via larger design margins) can minimize the failure rate during the ICs useful life. The third stage, i.e., the final servicing period, is characterized by a rapidly increasing failure rate, caused by accumulated wear-out (aging) of ICs' material/devices. To ensure wear-out occurs past the IC expected lifetime and not pre-maturely, and postpone its on-set as much as possible, preventive and corrective maintenance could be carried out starting from the beginning of the IC servicing life.

In time, as the technology nodes down-scale from micrometer range to nanometer range, devices are becoming less and less reliable, while the IC market demands products with higher reliability (e.g., increased yield, lower infant mortality, lower failure rate during the useful life stage of the "bathtub" curve, increased useful life span). With miniaturization, devices exhibit increasing variability in their characteristics w.r.t. nominal values, which can lead to large spreads in delay, power, and robustness, and make devices become more prone to variability and noise induced failures during their in-field usage, and it is in this landscape of increasing variability, that reliability targets must be met with viable costs.

Specifically, in the micrometer range, devices exhibit a relatively small variability of the critical physical and electrical parameters, which confer them the expected performance envelope. The traditional way for a commercial IC designer to deal with reliability concerns is to introduce higher, more conservative design guard bands (e.g., maximum operating frequency guard band of up to 20%), in order to account for the performance loss caused by in-field variability induced failures [14]. Hence, the commercial IC designer's effort and the associated area and delay overheads required to secure the targeted reliability figures of merit, are relatively small. On the contrary, for reliability critical ICs (with much more restrictive reliability targets), much more effort is involved, fault tolerant techniques being developed and in place since early times [15]. For long life, not maintained systems, dynamic recovery fault-tolerant techniques were being employed as early as the 1960's (e.g., the Self-Testing-And Repairing (STAR) computer developed at NASA for a 10-year space mission [16]). For ultra-dependable, real time applications (e.g., mass transportation, fly-by-wire avionics, nuclear power plants), fault masking via modular and hybrid redundancy (hardware replicas in different technology nodes; software fault tolerance) is commonly employed [17–19]. For systems that required high availability, such as those in telephone switching systems, or transactions processing, self-checking techniques on replicas are typically deployed [20].

As the technology aggressively downscales into the under 100 nanometer regime, as a result of various inaccuracies in the manufacturing line, of the inevitable electrical charge granularity and matter atomic scale (e.g., quantum-mechanical effects such as

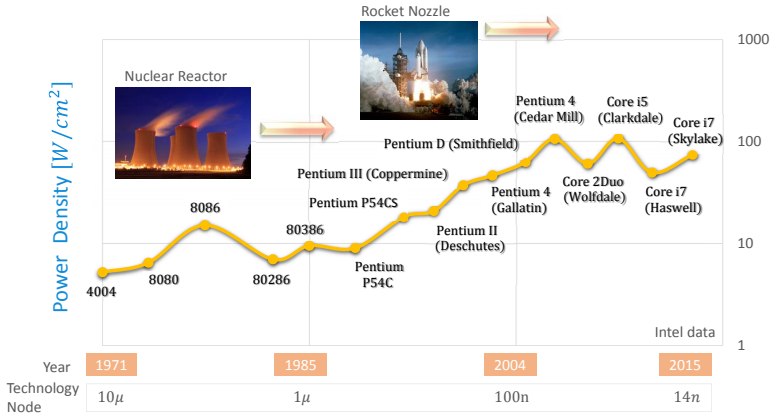


Figure 1.2: Power Density of Intel Microprocessors.

gate direct tunneling and energy quantization of carriers become visible), etc., devices exhibit increased variability of critical parameters, thus they are not any longer able to systematically deliver their nominal expected behavior. Moreover, as the number of transistors per area unit increases, the gate oxide electric field experiences a significant increase and the power density becomes significantly higher, as illustrated in Figure 1.2, leading to very high chip temperatures. Such high temperatures and electric fields, usually encountered during burn-in screening, are now being experienced by under 100nm commercial high performance ICs (besides the reliability critical ICs) during their useful life stage, and in normal operating conditions. Thus, with each new deca-nanometer technology node, the consequences of scaling are twofold: (i) the IC useful life is reduced, as the on-set of the final servicing life stage (i.e., the wear-out stage) is being accelerated, and (ii) the failure rate during the IC useful life is increased. This trend is graphically illustrated in Figure 1.1, by the sub 100nm red curve.

These two consequences descend from several variability sources affecting current nanoscale devices, whose taxonomy is presented in Figure 1.3. Device parameter variations can be broadly segregated into two coarse categories: spatial and temporal (lower half of Figure 1.3). The spatial process fluctuations of a device parameters, caused by manufacturing processes imperfections (manifested at time $t = 0$ of post-Si device lifetime), can be further subdivided into die-to-die variations (e.g., fluctuations of gate width (W), length (L), oxide thickness (T_{ox}), threshold voltage V_{th}, etc.) and within-die variations (e.g., random dopant concentration, line edge irregularities), both subcategories resulting in detrimental effects such as increased delay (mean and standard deviation), thermal runaways, and increased power and leakage spread. The temporal variations caused by temperature and voltage fluctuations, as well as wear-out intrinsic mechanisms such as Negative Bias Temperature Instability (NBTI), Hot Carriers Injection (HCI), Time Dependent Dielectric Breakdown (TDDB), Electro-Migration (EM), affect critical transistors parameters (e.g., thresh-

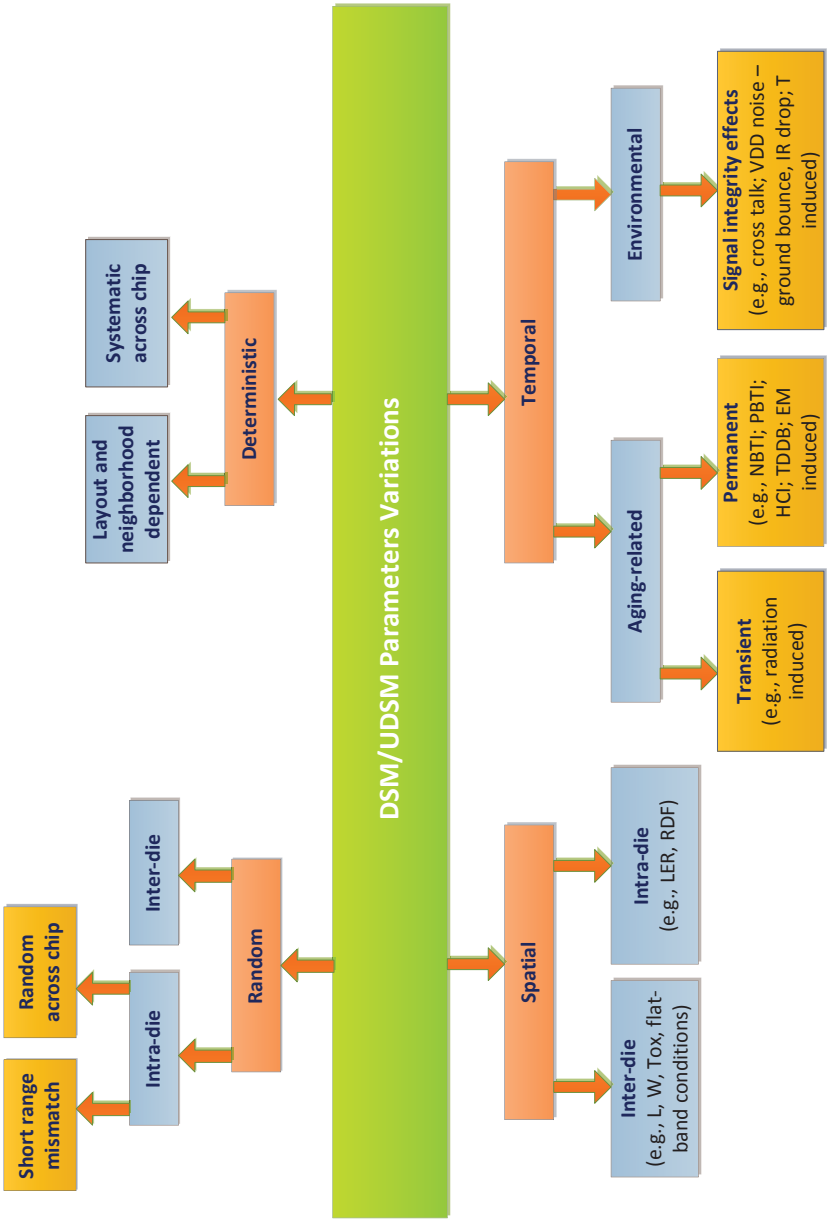


Figure 1.3: Structured View of the Different Types of Parameters Variations.

old voltage V_{th}, transconductance g_m, linear and saturation drain current I_D), induce abnormal delays and power dissipation, and shorten the device useful life. Another manner of categorizing variability is deterministic versus random, as depicted in the upper half of Figure 1.3. The systematic component has the advantage that is predictable, and once its influence on the transistor performance has been evaluated, it can be provisioned for in the design process, and thus completely eliminated. The random component on the other hand, can have its impact predicted only via a statistical characterization of the transistor/circuit behavior, and it is much more complex and costly to be accounted for.

As these variability sources have a higher impact on deca-nanometer than on micrometer range devices, ICs reliability targets cannot be any longer achieved solely by conservative design margins. On one hand, very large design margins would be required, which would impede attaining the maximal potential offered by the technology node and would significantly hurt performance and cost. On the other hand, due to faster device wear-out design guard bands might not be sufficient to ensure the lifetime reliability expectations. Therefore, for technology nodes $<$ 100nm, a significant threat to attaining the manufacturing yield with a viable cost and maintaining the reliability envelopes without placing a big burden on power and performance, is posed.

Neglecting the reliability concerns at design-time, is no longer a viable approach for a highly competitive semiconductor industry which emphasizes on short time-to-market, reduced Non-Recurring Engineering (NRE) costs associated with mask spins, first-pass success, and long-term reliability goals (e.g., extended useful lifetime). Specifically, reliability ought to be integrated into the design-time flow as an additional objective (besides area, delay, and power), circuit synthesis carried along such a multi-objective optimization setup, and reliability enhancing mechanisms providing the means for reaching the reliability targets during the IC intended lifetime, integrated within the circuit functionality. Moreover, in order to meet given in-filed demands, e.g., maximum failure rate, useful life length, reliability evaluation and mitigation issues should be also dealt with during the IC useful life "bathtub" curve segment. The reliability tasks performed in one phase of the IC lifetime, are often the result of the analysis and trade-offs performed in an earlier phase; thus a robust design constitutes a reliable IC foundation that enables an effective run-time lifetime management. Furthermore, the IC reliability has to be ensured via a closed-loop process, each phase providing feedback to previous phases to enable further reliability improvements (i.e., fewer encountered defects/failures) for the next generation ICs. To this end, the knowledge of the reliability profile/history over the complete IC life cycle, can serve to prevent failure recurrence by fixing its root cause, and not merely its symptom. In consequence, a deca-nanometer dependable IC needs an integrated approach addressing the reliability challenge both up-front, at design-time (pre-Si) and at run-time (post-Si). Most of existing efforts concentrate on dispersed reliability enhancing techniques, i.e., that are suited either solely for design time, or only for run-time, and most often without interfacing compatibility and interaction between design-time and run-time. However, to be effective, the required resiliency techniques for deca-nanometer ICs should transcend multiple levels of abstraction, including de-

vice, circuit, micro-architecture, architecture, and system, and envision cross-layer cooperation for optimizing the outcome.

In view of the above discussion we conclude that a framework able to deal with IC reliability aspects in an unitary manner, is crucial for the design and realization of dependable computing platforms and in this thesis, we propose an integrated framework aiming to address the reliability issues in a systematic way, from design to run-time.

1.2 Research Questions

In this section we introduce the basic principles governing the envisaged reliability aware design and life time management framework and formulate the main associated research questions addressed in this thesis. In a nutshell we propose a holistic framework, which systematically builds upon each abstraction level from device to system, and ensures inter-level operability in order to achieve a wear-out aware IC lifetime orchestration in line with user defined reliability targets and performance constraints.

The framework can be regarded as being composed out of two sub-frameworks, which inter-operate as follows: (i) the design-time sub-framework, which provides the reliability-aware adaptive architecture fabric, and (ii) the run-time sub-framework, which (a) dynamically manages the fabric wear-out profile, while fulfilling a set of user defined Quality-of-Service requirements (e.g., power lower than, throughput larger than, keep alive the key live-support components/tasks until, provide warnings if predicted time-to-failure is smaller than) based on information acquired by the fabric's network of sensors, and (b) maintain a full-life reliability log, to be fed-back to the design-time sub-framework in order to identify where most of the failures occurred and obtain hints about how to conduct the reliability aware design of the next IC generation.

The design-time sub-framework, schematically depicted in Figure 1.4, concerns itself with the reliability optimized and lifetime manageable hardware platform design, laying the infrastructure on which the run-time sub-framework operates. Specifically, its goals are: (i) perform a 4-dimensional Design Space Exploration (DSE), in order to obtain a reliability optimized circuit realization, that is compliant with given delay, area, and power constraints, (ii) pre-characterize the reliability enhanced circuit outputs Word Error Rate (WER) for a wide range of gate error probabilities, and (iii) generate a reliability wrapper containing the hardware means that allow the run-time sub-framework to observe and control the fabric according to the QoS specifications.

The DSE is concerned with the identification of a circuit realization able to perform the targeted computation, with a maximum output Word Error Rate (WER) α_C compliant with application/user defined reliability specifications, which could be much smaller than the targeted fabrication technology specific gate error rate α_G, during the circuit intended lifetime. More precisely, it is not an optimum circuit realization that is being sought, but rather a realization that fulfils the reliability constraints, while minimiz-

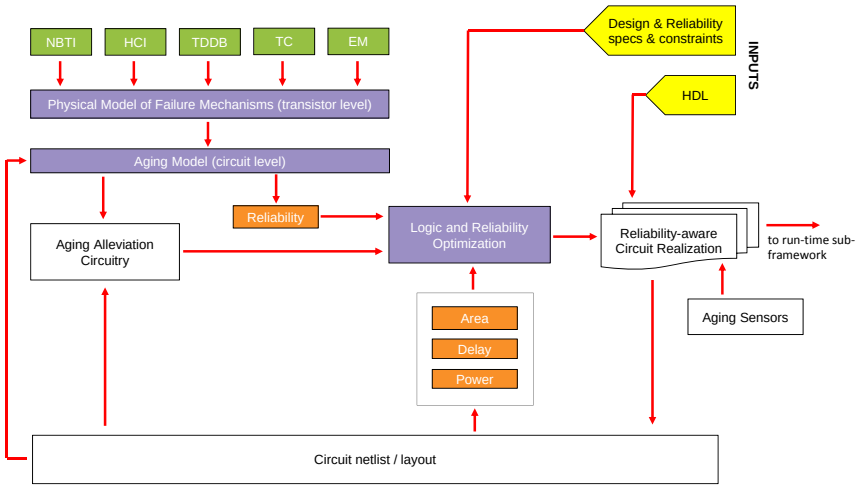


Figure 1.4: Reliability-Aware Circuit Design Sub-Framework.

ing the other 3 design constraints (i.e., area, delay, and power intent), possibly with different priorities. The design space exploration is performed via an iterative 2-step process: (i) conduct logic synthesis, and (ii) evaluate the performance and reliability of the circuit realization obtained from (i). To evaluate the reliability of a circuit realization, the circuit is subjected to an aggression profile that is likely to be encountered at run-time, and accelerated life simulation is performed (e.g., a device 10-year useful life is shrunk down to a very short period, such that the device reliability can be investigated and dealt with during that period). Aging models are then employed in order to infer the circuit reliability after Y years of operation, and predict its remaining useful life. If the output Word Error Rate (WER) and end-of-life targets are not being met, rewriting the initial circuit function, such that a more reliable circuit realization is obtained [21], and/or designing and employing reliability enhancers, e.g., modular redundancy [22], averaging cell [23] are pursued and the entire cycle repeated until an acceptable realization is identified.

Once the reliability optimized circuit is identified it has to be evaluated for gate error rates into a neighbourhood of α_G, to asses its output behaviour under various (other than the expected) aggression profiles. To this end a set of Monte Carlo simulations are required to estimate the circuit outputs WER in a 3-dimensional space, as a function of aging, temperature, and radiation, as they constitute the main sources of in-field degradation provisioned by sensors. The obtained WER surface is meant to serve as run-time reference for positioning the instantaneous WER (the WER point corresponding to the (aging, temperature, radiation) values currently sampled by the sensors), and as consequnece to pursue the action deemed as the most appropriate for that particular situation.

Given that in-field ICs may have to operate under different (harsher) aggression pro-

file than the one utilized during the DSE process the real wear-out after Y years of operation may be different than the expected one. Due to this, even though the circuit was designed to provide a smaller than α_C WER for its intended lifetime, it might fail to due so. To handle such situations the circuit has to be augmented with a reliability management wrapper, which creates the premises for lifetime adaptive, reliability-aware circuit management. The wrapper structure and detailed design depends on the circuit it protects but in principle it includes: (i) in-situ sensors (e.g., temperature, aging) for run-time fabric health status monitoring, (ii) mitigation/compensation mechanisms (e.g., Dynamic Frequency Voltage Scaling (DFVS), adaptive body biasing), (iii) adaptation knobs to control fabric operation regime according to the run-time sub-framework decisions, and (iv) a dedicated communication infrastructure to allow for sensor observation and knobs control.

Due to the high complexity of such an endeavour we do not target in this thesis the realization of the entire design-time sub-framework. Instead, we concentrate on its key aspects, i.e., reliability evaluation and reliability wrapper generation, as the viability of the proposed approach very much depends on the answers one can provide to the following related research questions.

In order to be able to assess the reliability of a circuit and guide the 4-dimensional (reliability, area, delay, power) DSE, models have to be devised that can project the degradation mechanisms physics to the transistor and circuit level electrical parameters. Extensive research effort - substantiated per se in the form of device parameters degradation models - was directed on understanding the physics and thermodynamics of the wear-out mechanisms [24–27]. A salient feature of existing device level aging models is that they capture the degradation dynamics of individual wear-out mechanisms [28–31]. In practice however, the devices are affected by multiple degradation mechanisms concomitantly, and their effects on the device parameters most often combine. This is also the case in the context of the proposed framework, thus the following research question arises:

- **Can we derive a device-level aging model that is able to capture in an accurate enough manner and with reasonable computational effort, the combined degradation effects of multiple, concomitant aging mechanisms?**

Such a model ought to be accurate enough from the aging assessment point of view, but also enable a fast enough computation, as circuits can contain thousands and hundreds of thousands of transistors. A fast reliability assessment benefits the design space exploration time required for reaching a circuit realization solution that is compliant with the circuit given performance and reliability specifications.

The next step is to determine which transistors are to be monitored for degradation, as we cannot monitor all transistors in a circuit, because of their large number. The common sense choice is to monitor only the transistors on the timing critical paths. However, a caveat of this approach is based on the observation that some transistors will degrade more than others, as a function of the input workload and environmental aggression factors they are being exposed to (e.g., temperature and supply voltage

fluctuations). The consequences are twofold: (i) a circuit path can become critical from the timing point of view, even if before it was not, as a result of its comprising transistors aging, and (ii) only a percentage of transistors on each path can significantly contribute to the circuit aging. This imposes a change of paradigm, more specifically, monitor only the transistors that could potentially cause the greatest circuit performance degradation due to their aging. Their numbers should also be as small as possible, as ultimately, the set of transistor chosen to be monitored at design-time determine the number and placement of aging sensors, which measure the amount of degradation at run-time, within the reliability wrapper. Identifying the transistors which influence the most the circuit aging and that ought to be monitored, and limiting their number, are two aspects which should be carefully considered and give shape to the following research question:

- **How to determine for a given circuit, the reduced set of transistors that are critical from the aging point of view and that should be monitored?**

Having determined the critical set of transistors in a circuit and their aging status, the next step in the reliability-aware design and computing framework is to determine the overall aging status of the circuit. As was the case also with existing transistor level aging models, most of existing circuit level aging assessment approaches, concentrate either on temporal variations [32, 33] or on process variations [34, 35], disregarding the interaction between them and their corroborated degradation effects on the overall circuit. We also note that previous approaches towards circuit level aging modeling are deterministic. While a deterministic approach is computationally faster, and better suited for larger circuits, a fully probabilistic one would be more accurate and appropriate for design-time, but also more complex. We note that an increased degree of reliability assessment accuracy is desired at design-time for two main reasons: (i) the effectiveness of the entire run-time reliability-aware management depends on the 4-dimensional DSE circuit realization outcome, and on the reliability alleviation mechanisms that are built-in at design-time, and (ii) if the aging assessment accounts for the history of parameters degradation, an inaccurate evaluation an time t could multiply its error magnitude at time $t+1$. The transistors parameters degradation can be either remnant or only temporary if the circuit was not exposed to stresses for a longer period of time, thus assessing only the instantaneous aging evaluation might be misleading. Regarding aging not only as a function of the instantaneous degradation of certain circuit parameters, but as a function of those parameters degradation evolution in time, ought to yield a more realistic circuit end-of-life prediction during the design-time. In view of the previous argument we propose to investigate the following questions:

- **For a given circuit topology, based on comprising transistors aging status, can we infer via fast deterministic means a circuit aging status, and predict its remaining useful life?**
- **Can we derive via fully probabilistic means a circuit aging status, such that its degradation history is taken into account?**

In order to satisfy user defined reliability requirements, for a given fabrication technology and under a certain run-time aggression profiles, thus to achieve reliable computing, the 4-dimensional DSM process may have to make use of reliability enhancing circuit modifications and fault tolerant techniques into the quest for the most appropriate circuit realization. Moreover, additional fault tolerant circuitry can be embedded into the reliability wrapper and only activated when due to extreme environmental aggression profile impede the circuit in delivering a smaller than α_C WER. Fault tolerant circuitry has to rely on a certain form of information, temporal, or spatial redundancy, (e.g., modular redundancy [22], averaging cell [23]), which while being effective from an error tolerance perspective, generally involves high area and/or performance impairments. Additionally, such techniques do have an Achilles' heel in the voting circuit and do not allow for fine grain tuning of the achieved WER vs. the overhead.

An alternative avenue is to investigate the potential Error Correction Codes (ECC) may have in enabling reliable computation with aged unreliable devices. This proved to be successful in protecting nanoscale memories where ECCs are frequently used to encode the to be stored data, allowing afterwards for the transient bit flips occurrences to be detected and corrected [36, 37]. For digital logic, however, the same line of thought is not directly applicable, as in this case it is not the data - the memory content - that is affected directly by errors, but it is the hardware that generates the data. The logic gates which perform the computation are affected by errors that propagate and cumulate until reaching the circuit primary outputs which are producing the to be protected data, which is raising the following research question:

- **Given a combinational logic circuit subjected to fault inducing conditions, can we augment the original circuit with an ECC codec able to protect the circuit primary outputs while being itself subjected to errors, such that after decoding, the correct, error-free original circuit outputs can be recovered?**

The aggressive technology scaling impacts memory and logic reliability but also the one of interconnects. Moreover, while the transistors performance improves with decreasing feature size, the wires performance does not, as when the feature size shrinks, wires are getting shorter, but also narrower and taller. Their resistance and capacitance per unit length increase, which translates to increased wire propagation delay and higher power consumption. In deca-nanometer technology nodes wire delay, especially global interconnects delay, has become a major limitation of high performance ICs timing characteristics, often being more dominant than the transistors switching delay. Additionally, due to their larger capacitance, interconnects have become the dominant source of power dissipation for contemporary IC designs, surpassing the gate and diffusion capacitances power [38]. Another harmful interconnects scaling consequence is that the capacitive and inductive coupling noise between adjacent wires is becoming more prominent and can severely impact the signals integrity and propagation delay. The technology scaling induced parameters variations, as exemplified in Figure 1.3, affect the reliability of data transport over wires, resulting in erro-

neous data transmission and contributing to a poorer performance of the interconnect and its signalling circuit. In order to deliver against the reliability and performance envelopes when transmitting data across a chip, reliability enhancing circuitry, generated by the design-time framework should be envisaged. To this end, we propose to investigate coding schemes that encode the data prior to its transmission over the wires, able to reduce the switching activity as well as the crosstalk-inducing coupling transitions between adjacent wires. A balance has to be found between codec complexity and the facilitated transmission energy reduction, such that the codec energy consumption does not nullify the transmission energy reduction. In this context, we pose the question:

- **Can we devise energy efficient and reliable data transport via an unreliable medium?**

At the crux of the reliability wrapper reside the aging sensors (and their associated communication infrastructure) as their accuracy and operating principle have an essential impact on the lifetime reliability management platform performance and associated overhead. Past approaches towards aging sensor designs [39–43], rely on indirect measurements, i.e., the to be monitored devices are replicated and then exposed to the same stress as the ones that part take into the computations, and typically employ as aging quantifier a low level device parameter, e.g., transistor threshold voltage V_{th}. These sensors use a performance comparison between stressed and fresh (unaged) devices in order to get the aging information. As a result, existing approaches to sensor designs, while having the advantage of non-intrusiveness (i.e., non-interfering with the circuit operation), rely on an indirect wear-out assessment, which by its very nature is less accurate than a direct one. Additionally, it would be desirable for the aging monitor to be an easier to measure electrical quantity (e.g., device voltage, current). This would decrease the complexity of the previous sensors approaches, which use as aging monitors transistor's internal parameters such as the threshold voltage, or the carriers mobility, which are not easily measurable. A direct measurement is also expected to improve the aging assessment quality, which is very important, as the effectiveness of the run-time reliability aware adopted policy greatly depends on the accuracy of the sensed aging data. The next research question thus formulates as:

- **Can we design cost effective aging sensors that provide a direct measurement of amalgamated wear-out effects?**

Thus far, we discussed the design-time sub-framework and the associated research questions we address in this thesis. We note that other design-time sub-framework related open issues exists, e.g., How to efficiently perform the 4-dimensional DSE?, but we concentrated on what we consider as being the enablers for the proposed framework.

As mentioned in the section opening, the reliability optimized circuit realization and the associated reliability wrapper generated by the design-time sub-framework constitute the foundation on which the run-time sub-framework, schematically depicted

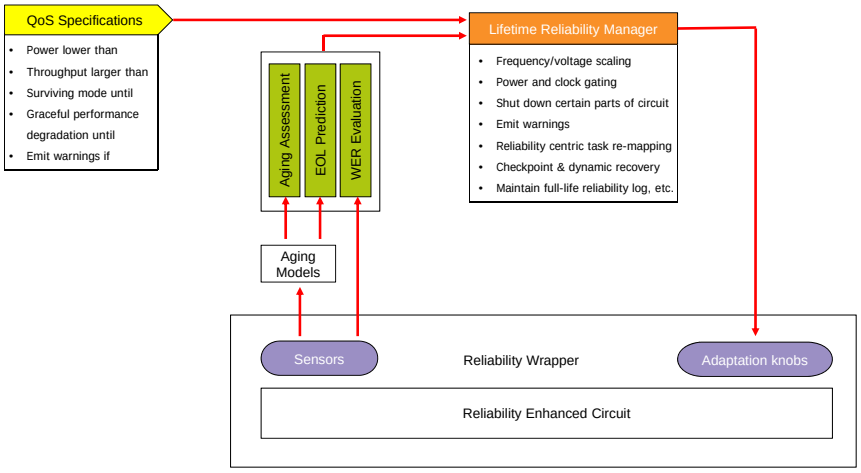


Figure 1.5: Dynamic Lifetime Resources Management Framework.

in Figure 1.5, operates. The run-time sub-framework concerns with two main functions: (i) evaluate circuit outputs WER and based on past and current raw sensors data, assesses the current circuit health and predicts its remaining useful life span, and (ii) reasons about adopting a particular reliability management strategy if the assessed circuit reliability is not compliant with the QoS specifications. We note that, in potential practical implementations the run-time sub-framework goes beyond reliability management only and deals with other QoS specifications, e.g., latency, throughput, power, too. Given that power and performance evaluation is out of this thesis scope we only discuss on the sequel the run-time sub-framework modus operandi solely from the reliability standpoint and note that it can in principle interoperate and potentially (partially) share infrastructure with state of the art performance targeted resource management platforms [44–46].

Specifically, the temperature and aging in-situ sensors that are part of the design-time reliability wrapper infrastructure sample in a quasi-continuous manner the circuit state. The collected raw sensor data are subsequently process by aging models to infer the circuit current aging status and predict its End-of-Life (EOL), based on current and possibly past sensor readings. Moreover, the actual circuit WER is assessed by identifying the circuit operation position, on the design-time pre-characterized WER surface, corresponding to the (aging, temperature, radiation) instantaneous values sampled by sensors. In this way we capture the actual position of the circuit within the reliability ”bathtub” framework both in terms of age and failure rate and derive the appropriate reliability management policy. In accordance with circuit reliability status compliance with the user-defined QoS requirements (e.g., WER smaller than, emit warnings if EOL earlier than, surviving mode when, graceful performance degradation when), the operation scenario (i.e., power supply, frequency, mitigation means) best-suited for the current circuit status is determined and put in place by means of

the adaptation knobs. For instance, if the delivered circuit outputs WER is bigger than the QoS specified acceptable error rate α_C, corrective actions which are changing the circuit operation mode (e.g., less workload, lower frequency, activate additional reliability circuitry) can be undertaken such that QoS acceptable WER figures are reached. On the contrary if WER is too low measures can be taken, e.g., power done reliability enhancers, in order to save energy while still fulfilling the QoS requirements.

If current wear-out is visibly reflected in the circuit behaviour such that (some) QoS constraints are violated, the voltage/frequency can be dynamically adapted in order to lessen the stress on the most affected parts of the circuit, or reliability-centric task re-mapping can be performed, such that less workload is distributed to the components which are most affected by aging. Resources which have pre-maturely aged, or permanently failed, can shift the course of action on surviving mode, i.e., only application life support functionality is provided by solely performing its essential tasks. Graceful performance degradation can come into play when certain circuit resources are either defective of highly faulty, in which case their tasks are relocated to other parts of the circuit which are functional, allowing the overall circuit to perform all expected tasks, but with a lower performance (e.g., slower, lower throughput, higher energy consumption). If the degradation corrective measures are not effective and certain degradation thresholds are being reached, warnings are emitted prompting for user intervention for further actions. Once a line of action has been determined, it is physically enforced via the design-time reliability wrapper knobs that control the circuitry responsible with e.g., voltage/frequency scaling, reliability enhancing circuits, e.g., TMR, and aging mitigation/compensation.

In this thesis, as previously mentioned, we only focus on the run-time framework part dedicated to the transformation of the raw data, acquired from aging and temperature sensors, into meaningful circuit/platform level wear-out information. This step is essential as the runtime reliability management paradigm decisions and actions builds upon its outcome. In order to assess the circuit aging and predict its remaining useful life span, the sensors information has to integrated via an aging model. Sensor measurements are typically noisy, as a result of various measurement imperfections, environmental disturbances, their own aging, etc. Existing literature assumes ideal measurement conditions [41–43] and do not take into account the impreciseness of the sensors measurements, which can lead to either circuit aging status overestimation or underestimation, and further to too restrictive, or less effective reliability mitigation/compensation measures. Thus, different from the design-time aging models, the run-time aging models ought to account for the uncertainty of the sensors measurements when deriving the overall circuit aging status and predict its end-of-life. Thus, the next research question to investigate can be formulated as follows:

- **How to integrate the noisy aging sensors measurements in a comprehensive manner in order to assess a circuit instantaneous aging status and predict its end-of-life?**

In answering the above 8 research questions, this thesis investigates reliability-aware design and dependable computing and data transport techniques in order to deliver

against the reliability envelopes of current technology nodes. Specifically, as presented in the next Section, we investigate, propose, and evaluate technique for (i) the design of reliability enhanced circuits and their enveloping reliability wrappers (sensors, communication infrastructure, and adaptive control mechanisms), and (ii) sensor data based reliability assessment for an effective in-field lifetime reliability management, in compliance with user defined QoS specifications.

1.3 Thesis Contributions

Subsequently, we summarize the contributions we made towards the potential realization of the proposed reliability aware design and life time management framework, as follows:

- We propose a design-time transistor level aging assessment and End-of-Life prediction model that uses the transistor output signal slope as aging quantifier, and can capture the amalgamated effects of the NBTI, HCI, and TDDB aging mechanisms. The model accounts for the intrinsic degradation of transistor's own parameters (e.g., V_{th}), which are ultimately reflected in the transistors's output voltage slope degradation. Moreover, it is also able to take into consideration the influence of the surrounding circuit topology via: (i) the slope degradation of the surrounding environment current contribution, which reflects the impact of the aging of neighbouring transistors connected to the source terminal of the transistor whose aging we assess, and (ii) the slope degradation of the transistor's gate voltage, which reflects the impact of the aging of the driver transistors. We define the transistor age as the time integral of the sum of output voltage slope variations caused by the aging induced drift of its intrinsic/extrinsic parameters. A NOR-2 gate implemented in TSMC 65nm CMOS technology is considered as test vehicle, and exposed to NBTI and HCI stress, for a temperature range of $25^\circ - 175^\circ C$. Compared against the Cadence simulation, we obtain a fairly small estimation error of the output voltage slope (mean $< 10\%$ and standard deviation $< 5\%$), which confirms the feasibility and potential of the proposed approach. Since PMOS transistors typically suffer more degradation than the NMOS counterparts, we consider the lower PMOS transistor in the NOR-2 gate, and evaluate the amount of degradation induced by: (i) only its intrinsic aging (which is the typical employed method), and (ii) both intrinsic and extrinsic (from the neighbours) aging, as proposed in our model. Simulation results reveal that if only intrinsic degradations are considered the PMOS transistor aging is underestimated by $4\% - 7\%$, when compared with the case when both the intrinsic and extrinsic aging are taken into account, which quantitatively proves that our model provides a better insight into the aging assessment issue.
- We introduce a model able to infer circuit aging based on the aging of com-

prising transistors. Specifically, we propose to predict the circuit End-Of-Life (EOL) as a sum of the critical transistors individual EOL values, weighted by circuit topology dependent coefficients that are fixed at design-time via a multi-response regression approach. Given that abstracting the circuit level aging status from the aging of all its component transistors is not an option for non-toy circuits we further propose a methodology to select a reduced set of circuit transistors, i.e., those which are critical from the aging point of view, based on whose degradation circuit aging assessment becomes computationally tractable. To this end, we identify the circuit paths which are critical from the aging point of view, i.e., the paths whose aging induced degraded performance violates their un-aged timing requirements. In view of the fact that from each such aging critical path, only a small percentage of transistors could potentially cause the greatest extent of damage we introduce a method to identify and select those as constituents of an aging critical transistors kernel. The transistors thus selected are the ones to have their aging monitored, and used to infer the aging of the overall circuit via the proposed topology dependent circuit-level model. When applied on the ISCAS-85 $c499$ circuit implemented in PTM 45nm CMOS technology our approach identified a 15 critical transistors kernel out of a total of 1526 transistors, which greatly benefits from the standpoints of wear-out measurements sites (the number of aging sensors) reduction and circuit aging derivation tractability. When inferring the circuit age with the proposed model based on the 15 identified critical transistors, we obtain a small circuit aging approximation error (mean $< 15\%$ and standard deviation $< 6\%$) w.r.t. the reference circuit aging obtained by means of Cadence simulation, which quantitatively confirms the evaluation accuracy of our approach. When reducing the number of sensors by $2/3$, i.e., only 5 transistors out of the kernel of 15 are utilized for aging evaluation, the EOL estimation accuracy is diminished with 18% (reported relative to the estimation accuracy achieved by employing the entire kernel of 15 critical transistors), which indicate that trade-offs between area overhead and EOL estimation accuracy are possible, the lifetime prediction being kept within reasonable bounds. The proposed circuit aging assessment approach is deterministic and fast, and requires a minimum number of in-field aging sensors (to monitor the circuit kernel of critical transistors), which makes it a good candidate for run-time aging evaluation.

- We propose a fully probabilistic circuit level aging model which regards age not only as a function of the instantaneous degradation, but also of past degradation. To this end we introduce a probabilistic model that allows the circuit performance (e.g., clock frequency) to vary in time, and has a Markovian character (the age being expressed as a function of the degradation parameters values at the current sampling moment and the ones from the previous sampling moments), as opposed to the traditional deterministic approach, which considers a fixed performance boundary and includes only the current sampling moment degradation values when inferring the age. In this way, we account

for both remnant and non-remnant degradation effects depending on the circuit stress duration. To deal with aging, a guard-band is typically selected at design time (i.e., the circuit maximum frequency is lowered to account for the worst case of performance loss due to aging). When employing our aging assessment method, the aging is estimated for expected typical circuit working conditions, allowing thus (i) a less conservative guard-band selection (a higher allowable circuit maximum operating frequency), and (ii) aging mitigation/compensation techniques optimized for the expected working conditions induced degradation. Simulation results performed for the ISCAS-85 $c499$ circuit in PTM 45nm CMOS technology (with a kernel of 15 critical transistors out of 1526 transistors) yield an approximation error with mean $< 10\%$ and standard deviation $< 15\%$ w.r.t. the counterpart values provided by Cadence simulations. As expected, the aging values obtained with our framework are further from the Cadence values, as we take into account more parameters and aging sources; in fact we are probably closer to the real EOL values (but we have no means to verify this conjecture). We note that the proposed Markovian aging model necessitates the monitoring of multiple degradation parameters per transistor, e.g., V_{th}, μ, and hence multiple sensors are required for one transistor, which makes this approach less feasible for run-time aging assessment and prediction, but better suited at design-time, enabling a robust, fast, and accurate aging evaluation.

- We design a novel, on-chip aging sensor that can capture the amalgamated effect of multiple degradation mechanisms, e.g., NBTI, HCI. We achieve this by detecting the degradation of the peak supply current of the circuit under observation, which is a signature of the total drain current and a function of multiple aging sensitive device parameters, such as threshold voltage and carriers mobility. The proposed sensor measures the peak current value and converts it into a pulse-width modulated signal, further used to extract the overall circuit aging status. It consists of: (i) a Current Peak Detector (CPD) which comprises a current memory cell for retaining the peak current value, and a current comparator to determine if the present supply current value is bigger than the stored peak value, in which case the peak current value in the memory cell is being updated, and (ii) a Current-to-Time (C2T) converter based on a thyristor delay element, which generates a pulse signal which length is proportional with the peak current value given by the CPD. Unlike existing aging sensors, which estimate the aging status indirectly by replicating devices and exposing the computig device and the monitor to the same stress, the proposed sensor allows for direct aging assessment for a single device and/or large circuit blocks. Furthermore, different from existing sensors which monitor the aging of individual transistors, the proposed sensor can monitor in principle the entire circuit which means thus that no transistor kernel identification and models for aging abstractization from transistor to circuit level are required in our case. We design the sensor in TSMC 65nm CMOS technology and conduct accelerated test simulations in Cadence for ISCAS-85 circuit benchmarks. Simulation results reveal

that the drain current exhibits a near identical aging rate as the threshold voltage V_{th}, for most of circuit lifetime. However, towards circuit EOL, the peak power supply current and the threshold voltage aging indications start to deviate from each other. Unfortunately, no other aging reference is available thus we cannot tell which method provides better estimates for that circuit lifetime part. We note however that high accuracy is essential during an IC early life, when it undergoes the greatest wearout induced degradation, demanding thus accurate and fast reliability assessment and EOL prediction for an appropriate degradation compensation/mitigation measure.

- We propose a run-time suited EOL and reliability estimation model, that can deal with the nonlinearities of the degradation process, and which takes into account the sensor measurements uncertainties. Once new degradation data are made available by the sensors, these data are employed together with the degradation history to update the aging estimation results via a Bayesian approach. For validating this model we use a power law degradation model. SPICE simulation results reveal that when compared to the commonly employed degradation process with linear mean, our approach yields improved estimation accuracy. The proposed model may thus be better equipped for capturing the nonlinearity and variability of in-field degradation dynamics, when compared to traditional models which assume ideal sensor measurements. We note that a better approximation accuracy is of particular interest especially for the circuit early-life stages (e.g., the first 2, or 3 years from a 10 years lifetime), because during this period the circuit suffers the greatest extent of wear-out, and as a result the more accurate the aging assessment, the more effective the aging mitigation/compensation techniques could be.

- We introduce and evaluate Haar based codec assisted medium and long range data transport structures (e.g., bus segments, Network on Chip interconnects), able to deal with technology scaling related phenomena (e.g., increased susceptibility to proximity coupling noise and transmission delay variability), targeting energy savings at the expense of a reasonably small overhead, i.e., 1 extra wire, a 2-gate encoder, and a 2-gate decoder, for each and every pair of uncoded wires. For practical evaluation we employ a commercial 45nm CMOS technology and different random, uncorrelated workload profiles. For 5mm and 10mm long 8-bit buses (without repeaters), we obtain energy savings of 55% and 34%, and a transmission frequency increase of 35% and 41%, respectively, at the expense of less than 1% area overhead with respect to the reference system (i.e., 8-wire synchronous uncoded bus), which prove energy and delay effectiveness. We further augment our proposal with a Single Error Correction and Double Error Detection (SECDED) scheme particularly adapted to its structure, in order to cope with very deep sub-micron noise (e.g., supply voltage variations, electromagnetic interference) induced transmission errors. When compared to the reference system (not SECDED protected), for 10mm long buses, the Haar tailored SECDED approach consumes 27% less energy at the expense of 2% area overhead, while providing robuster data transmission.

- The significant uncertainty associated with current nanodevices fabrication and operation, calls for a circuit design paradigm change, which ought to actively embrace the inherently nanodevice unreliability to generate overall circuit architectures able to perform reliable computation. While for data storage units viable Error Correcting Codes (ECCs) based solutions exist, Data Processing Units (DPUs) are not amenable to a similar line of reasoning. The typical approach undertaken for fault-tolerant DPUs relies on modular redundancy (e.g., spatial, temporal), which while being effective from the error reduction perspective, generally involves high area and/or performance impairments. To this end we propose a generic methodology to obtain reliable DPU implementations built with unreliable components by intimately intertwining ECCs codecs with the DPU functionality. The ECC protected DPU architecture is derived cluster-wise with area and reliability constraints, by exploiting dependence relations (logical and w.r.t. shared area) between internal signals pertaining to the DPU and the ECC codec. To evaluate the implications of our proposal in terms of error rate and performance, we consider a multitude of test corners (e.g., gate criticality, ECC type and structure, faulty and low complexity decoder, time-space redundancy) for an ECC protected 6-bit adder architecture, using TSMC 40nm CMOS technology. Simulation results reveal that the ECC embedding approach can be effective from both error rate and area perspective, for the Pareto designs with performance figures of merit situated in-between consecutive modular redundancy based design corresponding curves. For instance, the adder protected with an optimal code w.r.t. code length and with an error correction capacity of 5, has an area overhead of 33% relative to the 3-modular redundancy design, but with 20% less than the 5-modular redundancy design, while from the error rate point of view it is situated approximately in the middle of the error rate characteristics of the two modular redundant designs. Additionally, our approach eliminates the modular redundancy Achilles' heel as the centralised majority voting is decentralized in our case, which potentially results in robuster circuit implementations. The proposed approach is generic from the coding point of view, scalable, and enables a fine grained control of the DPU specifications levels of reliability and maximum area overhead.

1.4 Thesis Organization

The remainder of the thesis is structured in 7 chapters, as follows:

In **Chapter 2**, we present a transistor-level aging assessment model, that uses the signal slope as aging quantifier and captures the joint effect of multiple concomitant aging mechanisms. To improve the aging assessment accuracy, the model takes into account besides the conventional device intrinsic self-degradation, the influence of the surrounding degradation affected circuit topology.

In **Chapter 3**, we propose an aging assessment model, which integrates at the circuit-level the transistor-level aging information. A selection methodology of a minimum size kernel of aging critical transistors is introduced. Based on the aging information provided by the aging critical transistors, two methods for estimating a circuit End-Of-Life (EOL) are proposed: (i) expressing the circuit EOL as a sum of the critical transistors EOL values, weighted by fixed, topology dependent coefficients (deterministic method), and (ii) deriving the circuit EOL via a statistical Markovian model that takes into account the degradation history and accommodates for both remnant and non-remnant degradation effects (probabilistic method).

Chapter 4 is concerned with the design of an on-chip aging sensor able to capture the amalgamated aging effect of multiple degradation mechanisms, e.g., NBTI, HCI, on large circuit blocks. The proposed aging sensor monitors the peak power supply current degradation, which serves as aging quantifier. The sensor allows for direct measurement of device or even larger circuit blocks degradation, unlike existing sensors which can only monitor devices and get indirect measurements from their replicas.

In **Chapter 5**, we present a statistical Bayesian framework to estimate circuit's EOL and reliability status, based on in-filed aging sensor measurements. The framework accounts for the nonlinearities of the circuit degradation process, as well as for aging sensors measurements imprecision.

In **Chapter 6**, we introduce and evaluate codec assisted data transport structures, suitable for medium/long range interconnects, able to diminish the energy consumption via a Haar transform and cope with transmission errors via a Haar tailored Single Error Correction and Double Error Detection mechanism.

In **Chapter 7**, we concern with reliable data processing and investigate whether the avenue of functional unit reliability improvement via intimately intertwining the original functional unit circuit and an Error Correcting Code (ECC) codec, could prove efficient. To this end, we proposed a framework to derive the ECC extended architecture of a given functional unit, and perform extensive design space exploration, spanning various architectural and ECC related test corners.

Chapter 8 summarizes the thesis accomplishments, and provides possible directions for further pursuing this line of research.

2

Design-Time Transistor-Level Aging Model

Accurate age modeling, and fast, yet robust reliability sign-off emerged as mandatory IC design constraints for advanced process technology nodes. Reliability evaluation at design-time is of interest, as it enables the identification of circuit architectures that take reliability into account, besides the traditional delay, power, and area objectives. In this chapter, we propose a design-time device-level aging assessment and prediction model that makes use of the transistor output signal slope as aging quantifier, and accounts not only for the intrinsic self-degradation but also for the influence of the surrounding circuit topology. The model is able to capture the joint effect of multiple concomitant aging mechanisms, e.g., NBTI, HCI. Experimental results indicate the validity of slope as aging quantifier and that aging is underestimated when topology influence is disregarded.

2.1 Introduction

Wear out mechanisms, further enhanced by the aggressive CMOS scaling adopted for performance improvement, have emerged as major reliability concerns of deep sub-micron devices. These aging phenomena lead to time dependent degradation of critical physical and electrical transistor parameters, which degrades the performance and eventually produces device failure. These considerations in addition with the high pressure in achieving short time-to-market figures imposed the reliability analysis to be addressed earlier in the product fabrication cycle, namely at design time.

Existing device-level reliability assessment approaches use delay as aging monitor [47–49], and model it as function of a parameter drift, obtained either analytically or by means of simulations. The delay models are commonly based on the Sakurai's α power law MOSFET model, [50], to express the transistor current, which does not take into account the prevalent effects characteristic for present nanometer technologies. Moreover, most papers do not model the joint effect of multiple aging mecha-

nisms [28–31] and very few consider the aged signal slope as wearout monitor [51]. In [52], the authors propose a delay model considering both Time Dependent Dielectric Breakdown (TDDB) and Negative Bias Temperature Instability (NBTI). In [53] an NBTI and Hot Carrier Injection (HCI) aware delay model that also provides the degradation of the output slope is introduced.

In this chapter we propose a general framework for measuring the age of a device characterized by the fact that the parameter used as an aging measure is merely a functional of the aging parameters as functions of time. We propose to use as example of aging quantifier the signal slope at half the supply voltage, i.e., at $V_{DD}/2$. We describe the slope based aging model analytically and validate it experimentally. The model takes into account the influences of the neighboring devices as well as own device degradation caused by Front-End-Of-Line (FEOL) aging mechanisms. Experimental results indicate that the aging prediction accuracy is improved when compared to the case when the external topology influence is disregarded.

The chapter is organized as follows: In Section 2.2 we introduce the signal slope as aging monitor. Section 2.3 provides an explicit description of the proposed level aging model. The model accuracy is then evaluated in Section 2.4. Finally, Section 2.5 provides a summary and comments on the results.

2.2 Signal Slope as Aging Monitor

The circuit employed for device age characterization is graphically depicted in Figure 2.1, where:

- V_{in} denotes the gate voltage, with slope P_{in},
- V_{out} denotes the output voltage, with slope P_{out}, and
- I_x is the surrounding environment current contribution, with slope P_x.

Also, C_M and C_L are the input to output coupling capacitance and the output capacitive load, respectively. For proper transistor lifetime characterization, one should take

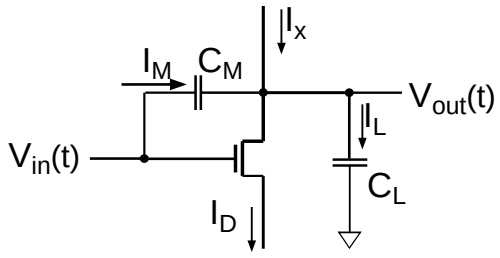


Figure 2.1: Transistor Age Assessment Circuit Schematic.

into account not only the intrinsic self-degradation, but also the influence of topology on the transistor in question (i.e., the influence of degraded transistors adjacent to the transistor under study). We propose to account for this influence of adjacent transistors, by means of:

(i) variation of gate voltage slope $\Delta dV_{in}/dt$ - the impact of aging of driver transistors; and

(ii) variation of I_x current slope $\Delta dI_x/dt$ - the impact of aging of transistors connected to the source terminal.

The modification of the voltage slope from device input to its output, namely: $P_\Delta = P_{in} - P_{out}$, measures the influence of the device in the degradation of the signal. We express P_Δ as being composed of two terms: $P_\Delta = P_{\Delta a} + P_{\Delta 0}$, where $P_{\Delta 0}$ accounts for the inherent and initial degradation of the slope; and $P_{\Delta a}$ is the part of the degradation which increases with the device age. In principle, the term $P_{\Delta a}$ accounts for all factors which negatively impact the device performance, that is:

(i) intrinsic factors - the drift of own degradation parameters $\mathbf{X}$ (e.g., V_{th}), and

(ii) extrinsic factors - the variation of P_{in} and P_x slopes.

More formally, $P_{\Delta a}$ can be expressed as a functional as follows:

$$P_{\Delta a} = f(\mathbf{X}(.),\ P_{in}(.),\ P_x(.)).$$

In view of the above, the age can be defined through the time integral $\int_0^t dP_{\Delta a} dt$, where $dP_{\Delta a} = dA$ are the time decrements of the slope. Once the aging increment is computed, one can proceed with the derivation of the aging rate and age expressions, i.e., the aging rate is derived by taking the ratio between the aging and time increments $(A_{rate} = dP_{\Delta a}\ /\ dt)$ and the age is given by integrating the aging increment over the interval $[0,\ t]$. In the analysis of a single transistor one can consider $P_{in} =$ constant (i.e., the input signal is always not degraded) and in general we normalize the age such that $A = k \cdot (P_{in} - P_{out})$ arrives at the value 1 when P_{out} arrives at 0.9 of its initial value P_{out0} for a given standard value of P_{in}. In consequence, for estimating the age of a transistor in real operating conditions we have to compute the value:

$$A = k \cdot \int_0^t (dP_{in} - dP_{out})dt = k \cdot \int_0^t A_{rate} dt.$$

2.3 Transistor Aging Model

The differential equation that describes the dynamic behavior of the circuit depicted in Figure 2.1, is derived by applying Kirchoff current law at the output node as follows:

$$\begin{aligned} I_x &= I_D + I_L - I_M \\ I_x &= I_D + (C_L + C_M) \cdot V'_{out}(t) - C_M \cdot V'_{in}(t). \end{aligned} \tag{2.1}$$

We assume that the transient analysis is concerned with the behavior of the NMOS transistor when the input voltage ramp V_{in} is rising, thus when C_L is discharging.

From (2.1) one gets immediately:

$$I'_x(t) = I'_D(t) + (C_M + C_L) \cdot V''_{out}(t) - C_M \cdot V''_{in}(t). \tag{2.2}$$

Consequently, we deduce the analytic expressions of the terms involved in (2.2), as functions of $V_{out}(t)$ and its time derivatives.

We adopt the ultra-compact I-V model from [54], which uses 9 parameters to characterize the transistor behavior in superthreshold conduction. Our choice is motivated by the fact that the model is simple (linear model) and accurate (average absolute error $< 10\%$ when compared to the state-of-the-art quadratic model derived by the same authors), while accounting for the prevalent physical effects encountered in current nanometer technologies. The current I_D is a function of t through V_{in} and V_{out}, so we can write:

$$I'_D(t) = \frac{\partial I_D}{\partial V_{in}(t)} \cdot V'_{in}(t) + \frac{\partial I_D}{\partial V_{out}(t)} \cdot V'_{out}(t). \tag{2.3}$$

Since for short-channel MOSFET devices, the carriers' velocity can saturate before the pinch-off condition is attained (L short enough so that $V_{DSsat} << V_{GS} - V_{TH}$), we consider the transistor operating in the saturation region at halfway of the output switching cycle ($V_{out} = V_{DD}/2$). Accordingly, we have the following expressions for the derivatives of the velocity-saturated drain current, with respect to the input and output voltages:

$$\frac{\partial I_D}{\partial V_{in}} = \beta_1 \cdot V_{out} + \beta_2; \quad \frac{\partial I_D}{\partial V_{out}} = \beta_3, \tag{2.4}$$

where the analytic expressions of β_1, β_2, and β_3 are derived subsequently.

Given the expression of I_D, that is:

$$I_D = P(a \cdot V_{out} + b),$$

the following partial derivatives result:

$$\begin{aligned}
\frac{\partial I_D}{\partial V_{in}} &= V_{out} \cdot \underbrace{(a' \cdot P - w \cdot k_{RDS} \cdot a' \cdot a)}_{\beta_1} + \underbrace{(P \cdot b' - w \cdot k_{RDS} \cdot a' \cdot b)}_{\beta_2} = \\
&= V_{out} \cdot \beta_1 + \beta_2 \\
\frac{\partial I_D}{\partial V_{out}} &= P \cdot a = \beta_3,
\end{aligned}$$

where:

$$
\begin{aligned}
P &= (w + k_{NWE} - w \cdot k_{RDS} \cdot a),\\
a &= k_I \cdot k_{DIBL} \cdot \frac{(V_{in} - V_{th})^{\alpha-1}}{V_{th0}^{\alpha}} + k_I \cdot k_{CLM} \cdot \frac{(V_{in} - V_{th})^{\alpha}}{V_{th0}^{\alpha}},\\
a' &= k_I \cdot k_{DIBL} \cdot (\alpha - 1) \cdot \frac{(V_{in} - V_{th})^{\alpha-2}}{V_{th0}^{\alpha}} + k_I \cdot k_{CLM} \cdot \alpha \cdot \frac{(V_{in} - V_{th})^{\alpha-1}}{V_{th0}^{\alpha}},\\
b &= k_I \cdot \frac{(V_{in} - V_{th})^{\alpha}}{V_{th0}^{\alpha}} - k_I \cdot k_V \cdot k_{DIBL} \cdot \frac{(V_{in} - V_{th})^{3\alpha/2-1}}{V_{th0}^{3\alpha/2}} -\\
&\quad - k_I \cdot k_V \cdot k_{CLM} \cdot \frac{(V_{in} - V_{th})^{3\alpha/2}}{V_{th0}^{3\alpha/2}},\\
b' &= k_I \cdot \alpha \cdot \frac{(V_{in} - V_{th})^{\alpha-1}}{V_{th0}^{\alpha}} - k_I \cdot k_V \cdot k_{DIBL} \cdot \left(\frac{3\alpha}{2} - 1\right) \cdot \frac{(V_{in} - V_{th})^{3\alpha/2-2}}{V_{th0}^{3\alpha/2}}\\
&\quad - k_I \cdot k_V \cdot k_{CLM} \cdot \frac{3\alpha}{2} \cdot \frac{(V_{in} - V_{th})^{3\alpha/2-1}}{V_{th0}^{3\alpha/2}}.
\end{aligned}
$$

The terms/factors k_I, k_V, k_{NWE}, k_{RDS}, k_{DIBL}, k_{CLM}, α, V_{th0} are the transistor ultra-compact model parameters [54] associated with the prevalent nanometer CMOS transistor physical effects, such as Short Channel Effects (SCE), Narrow Width Effects (NWE), Mobility Degradation (MD), Velocity Saturation (VS), Drain-Induced Barrier Lowering (DIBL), Channel Length Modulation (CLM), and Source-Drain parasitic Resistance (RDS).

Having derived the expressions of β_1, β_2, and β_3, it follows from (2.4) that:

$$I'_D(t) = \beta_1 \cdot P_{in} \cdot V_{out} + \beta_3 \cdot V'_{out}(t) + P_{in} \cdot \beta_2.$$

Finally, (2.2) becomes:

$$V''_{out}(t) \cdot (C_M + C_L) + V'_{out}(t) \cdot \beta_3 + V_{out}(t) \cdot P_{in} \cdot \beta_1 = P_x - P_{in} \cdot \beta_2, \qquad (2.5)$$

which is a second order linear differential equation with variable coefficients. We simplify (2.5) by using constant values for the coefficients (the values they have when V_{out} reaches $V_{DD}/2$). This is motivated by the fact that the equation is valid for these values for the linear domain of the commutation, i.e., in the neighborhood of $V_{DD}/2$ even if not always for the entire range (from the beginning till the end of commutation process). Nevertheless, for some initial conditions like $V_{out}(0) = V_{DD}$ and $V'_{out}(0) = 0$, one can extend the approximate validity till the moment when $V_{out}(t) = V_{DD}/2$, with a small approximation error on the derivative P_{out}.

Substituting V_{out} with e^{mt} and solving (2.5) for m, yields two solutions:

$$m_{1,2} = \frac{-\beta_3 \pm \sqrt{\beta_3^2 - 4(C_M + C_L)\,[P_{in} \cdot (\beta_1 + \beta_2) - P_x]}}{2(C_M + C_L)}.$$

Table 2.1: The Output Voltage Value, V_{out}, Sampled at Two Different Time Instants ($1/m$ and $2/m$), for Three Different Multiplicity Factors k (3, 1.5, and 1.1).

k	$V_{out}(1/m)$	$V_{out}(2/m)$	$V_{out}(\tau)$
3	$1.054 \cdot \frac{V_{DD}}{2}$	$0.4036 \cdot \frac{V_{DD}}{2}$	$0.9769 \cdot \frac{V_{DD}}{2}$
1.5	$1.315 \cdot \frac{V_{DD}}{2}$	$0.6123 \cdot \frac{V_{DD}}{2}$	$0.9542 \cdot \frac{V_{DD}}{2}$
1.1	$1.435 \cdot \frac{V_{DD}}{2}$	$0.761 \cdot \frac{V_{DD}}{2}$	$0.9713 \cdot \frac{V_{DD}}{2}$

The general solution of the differential equation (2.5) has the form:

$$V_{out}(t) = \mathcal{C}_1 \cdot e^{m_1 \cdot t} + \mathcal{C}_2 \cdot e^{m_2 \cdot t}, \tag{2.6}$$

where the constants $\mathcal{C}_1$ and $\mathcal{C}_2$ are obtained such that they satisfy the initial conditions corresponding to the transistor static operating regime, that is $V_{out}(0) = V_{DD}$ and $V'_{out}(0) = 0$. For usual values of the positive parameters β_1, β_2, and β_3, we may encounter the following two situations: (i) two real negative solutions, and (ii) two complex conjugated solutions (the third case with two real solutions of different signs was not encountered in practical simulations and thus it is not discussed in this chapter). In the following we detail each of these two situations.
Let us consider first that (2.6) has two real, negative solutions (i.e., $m_1 = -m$ and $m_2 = -k \cdot m$, where $k > 1$).

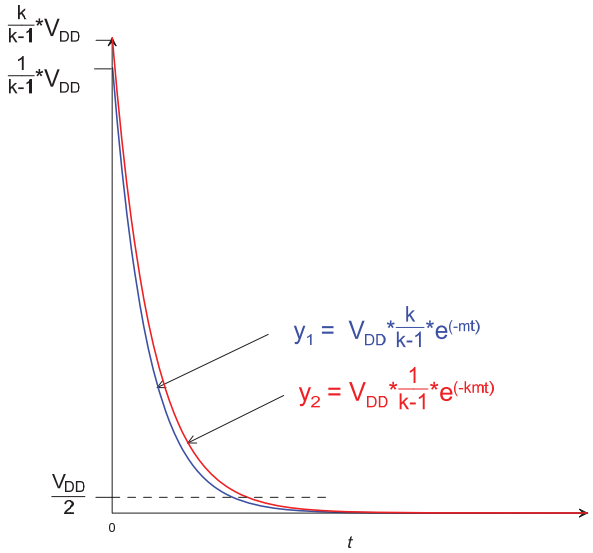
By appropriate substitution of $m_{1,2}$ and $\mathcal{C}_{1,2}$ as functions of m and k, we can rewrite (2.6) as:

$$V_{out}(t) = V_{DD} \frac{k}{(k-1)} \cdot e^{-m \cdot t} - V_{DD} \frac{1}{(k-1)} \cdot e^{-km \cdot t}.$$

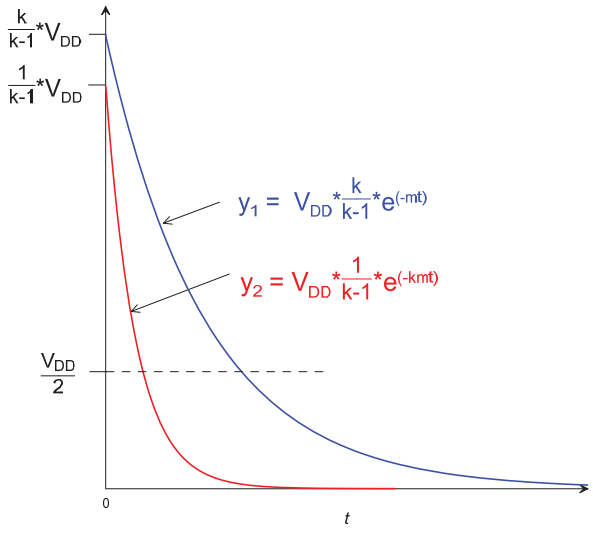
From here one can easily get the expressions for $V_{out}(1/m)$ and $V_{out}(2/m)$.

As one can observe in Table 2.1, the output voltage $V_{DD}/2$ is obtained between two time instants $1/m$ and $2/m$, for all considered values of k. Since $m_2 < m_1$, the exponential $e^{m_2 t}$ decays more rapidly than the exponential $e^{m_1 t}$. In consequence, for large values of k ($k \geq 3$), V_{out} is dominated by the first exponential contribution, while for small values of k ($1 < k < 1.1$), V_{out} is obtained by taking the difference of the two exponentials, both with similarly high amplitude. These two cases are graphically illustrated in Figures 2.2(a) and 2.2(b).

Based on the above observations, we conclude that the value of t for which $V_{out} = V_{DD}/2$ can be determined in all cases (under the assumption of two real negative



(a) small values of k, e.g., $k < 1.1$.



(b) big values of k, e.g., $k > 3$.

Figure 2.2: Graphical Interpretation of the Approximation τ.

solutions $m_{1,2}$), and propose as a good enough approximation of this t, the value found by the linear interpolation of the V_{out} values at $1/m$ and $2/m$ time instants. Explicitly, the searched t can be obtained by:

$$\tau = t_{linterp} = \frac{1}{m} \cdot \left[1 + \frac{V_{out}(1/m) - V_{DD}/2}{V_{out}(1/m) - V_{out}(2/m)}\right].$$

To confirm the assertion that approximation for τ is consistent, in Table 2.1, column 4, we evaluate $V_{out}(\tau)$ for the three considered k values.

Having τ, the slope V'_{out} at $V_{DD}/2$ becomes:

$$P_{out} = V'_{out}(\tau) = V_{DD} \cdot \frac{k\,m}{k-1}\left[e^{-km\cdot\tau} - e^{-m\cdot\tau}\right].$$

As concerns the solution of (2.5) for the set of complex conjugate values of $m_1 = m_r + j \cdot m_i$ and $m_2 = m_r - j \cdot m_i$, we obtain the following expression:

$$V_{out} = V_{DD} \cdot e^{m_r \cdot t}\left[\cos(m_i \cdot t) - \frac{m_r}{m_i} \cdot \sin(m_i \cdot t)\right].$$

Once we get a formula for P_{out} in which are included aging effects like the evolution of $\mathbf{Y}$ (e.g., V_{th}, P_{in}, P_x) with age, we can use it to derive the evolution of P_{out}. The following remark is in order: as soon as all the dependencies are of the type we deduced above, we can measure the age by the slope or deduce it from the values of the degradation parameters $\mathbf{Y}$. A more correct point of view is to consider, as mentioned in Section 2.2, that the slope is not a function of the instantaneous values of its input variables in a functional formula. It is a function of the entire evolution of those quantities and only its increase, i.e., dP_{out} is depending on the instantaneous values of variables and so $dP_{out} = \sum \partial P_{out}/\partial Y_i$ and

$$P_{out} = \int_0^t \sum \left(\frac{\partial P_{out}}{\partial Y_i} \cdot \frac{\partial Y_i}{\partial t}\right) dt.$$

All said about P_{out} could be translated for other measures and one of interest is the delay of a pulse produced by the aging. There could be situations when the slope has very small variations while the delay can be more easily measured and thus can serve as a measure of aging too.

2.4 Performance Evaluation

All simulations are performed in Cadence Virtuoso 6.1.5 [55], using a commercial 65nm Low Power (LP) technology. The reliability analysis is carried out using Cadence RelXpert and Virtuoso Spectre simulators. The following two modules are implemented in Verilog-A HDL: (i) the CMOS ultra-compact model from [54], with the

required 9 parameters extracted for bsim4 [56], and (ii) the aging module that consists of the equations that approximate the output slope and the aging equations. For the purpose of illustration, we consider V_{th} as the degradation parameter X, whose drift is caused by NBTI and HCI aging mechanisms [27].

In order to validate our model and evaluate its aging assessment and prediction accuracy, we employ as test vehicle a CMOS 2-input NOR gate, as depicted in Figure 2.3. For model validation, we use as devices under test the PMOS transistor M_3 and the

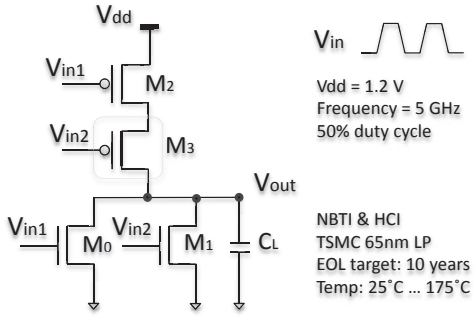


Figure 2.3: Test Circuit for Model Validation and Aging Assessment and Prediction Accuracy.

NMOS transistor M_1, both of them fresh (un-aged). The capacitance C_L is computed as the sum of drain junction capacitances and gate capacitances of fanout devices; the gate-drain capacitive coupling C_M is obtained as the sum of the gate-to-drain/source overlap capacitance and a fraction of the gate-to-channel capacitance. We apply various slew rates P_{in} and P_x and output loads such that both cases for the values of the 2 solutions of (2.5) are encountered.

In Figure 2.4, we compare for the PMOS transistor M_3 the estimated output slope at $V_{DD}/2$ using our model (red) with the one obtained from Cadence simulation (blue). By inspection of Figure 2.4, we note that in spite of the rough approximation quality of (2.5) on the whole time domain, the slope is approximated fairly well at $V_{DD}/2$. This is quantitatively confirmed by the relatively small values of mean ($< 10\%$) and standard deviation ($< 5\%$) for the approximation error.

Commercial aging modeling and reliability analysis tools, e.g., Eldo from Mentor Graphics, RelXpert from Cadence, do not integrate the influence of neighbor degraded devices when computing the age of the current device, i.e., a transistor exhibits during its lifetime an aging rate that is the same in the case when its adjacent neighbor transistor is fresh, as well as in the case when the neighbor transistor is aged. The total electrical stress experienced by a device during its lifetime is determined by the integration of the device stress rate during a transient simulation, followed by the extrapolation of the result at a specified age, e.g., at the end of the device lifetime. The device degraded parameters drifts are then obtained as functions of the total stress.

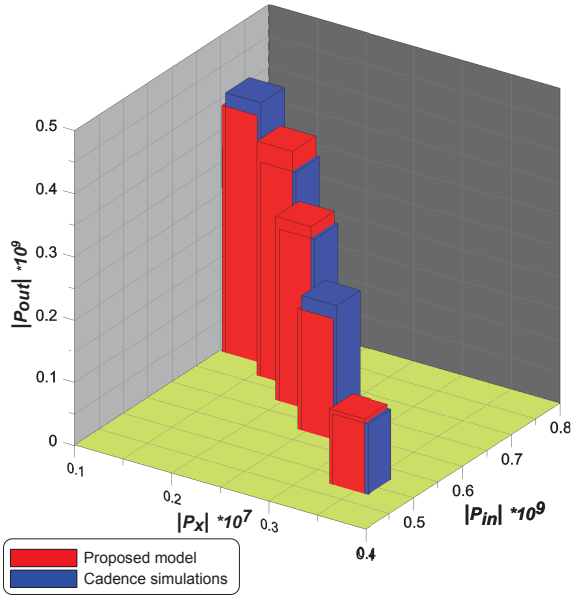


Figure 2.4: Simulated vs. Estimated Output Voltage Slope.

For deriving the stress rate are used accelerated lifetime models for the intrinsic FEOL mechanisms. Not taking into account the extrinsic contribution from the aged devices, may yield misleading results in terms of aging modeling and characterization.

In order to evaluate the accuracy of our slope-based aging model that accounts for both the intrinsic and extrinsic degradation, we compute M_3 transistor degradation in two cases: (i) with intrinsic and extrinsic influence factors (context-aware) and (ii) only with intrinsic influence factors. The corresponding simulation results are presented in Table 2.2, where $\Delta D = D_2 - D_1$, with D_1 [%] and D_2 [%] representing P_{out} degradation after 10 years for the context-aware case and for the intrinsic case, respectively. We apply several input aggression profiles typical for NBTI (static operation state) and HCI (dynamic switching periods), for a temperature range of $25°C$ - $175°C$, with increments of $50°C$. We adopt the usually encountered End-Of-Life (EOL) target to estimate the NBTI and HCI sensitivity to the CMOS process, that is 10 years. Examining Table 2.2, we note that when the aging assessment is performed disregarding the influence of aging of the neighbor transistor M_2, the amount of degradation of transistor M_3 is underestimated by $4\% - 7\%$ (for the considered simulation setup) when compared to the case when both the intrinsic and the extrinsic degradation factors are taken into account.

Table 2.2: Aging Assessment and Prediction Accuracy Context-Aware vs. Intrinsic.

	ΔD [%] @ M_3 after 10 years			
Aggression profiles	$25°C$	$75°C$	$125°C$	$175°C$
$V_{in1} = 0$; $V_{in2} = V_{in}$	-4.78	-6.22	-6.63	-6.90
$V_{in1} = V_{in}$; $V_{in2} = 0$	-4.98	-5.91	-6.32	-6.67
$V_{in1} = V_{in}$; $V_{in2} = V_{in}$	-4.23	-5.56	-5.95	-6.37

2.5 Conclusion

The temporal degradation of transistor performance is a key limiting factor of a circuit lifetime. As a consequence, aging assessment becomes of foremost importance for reliable deep sub-micrometer CMOS devices. In this chapter, we proposed a transistor level aging model that takes into consideration not only the own device degradation caused by FEOL aging mechanisms such as NBTI, HCI, and TDDB, but also the influence of the topology (i.e., the influence of aged neighbor transistors on the transistor whose age is being assessed). For this purpose, we defined the transistor age as the time integral of the sum of output voltage slope variations caused by the aging induced drift of intrinsic/extrinsic parameters (e.g., threshold voltage V_{th}, gate voltage slope P_{in}, output voltage slope P_x). The experimental results indicated that the slope of a transistor output voltage proves to be a reliable measure of aging able to capture the global result of: (i) own degradation due to working regime and environmental conditions and (ii) aging of other devices which influence the operation of the transistor under study. A final remark about the rational of this investigation is in order: there are efficient solutions to monitor the slope of the signal in a few points of a large circuit and so to monitor the circuit age; studies like ours prove that simpler models could help having a correct insight to the problem.

In the next chapter we address aging modeling at the next abstraction level and, by making use of the transistor-level aging model, we first propose a modality to identify the location and number of the to be monitored transistors from which status we can extrapolate the degradation from device to circuit level. Subsequently, based on the aging status of the transistors identified for aging monitoring, we introduce a design-time circuit level aging model.

Note. The content of this chapter is based on the following paper:

N. Cucu Laurenciu, S. D. Cotofana, **Context Aware Slope Based Transistor-Level Aging Model**, Microelectronics Reliability 52(9–10), pp. 1791–1796 (2012).

3

Design-Time Transistor-Based Circuit-Level Aging Model

In this chapter we introduce a novel method to assess circuit reliability and predict its End-Of-Life (EOL) at design time as well as at run-time. The main goal of our proposal it to allow for: (i) design time reliability optimization, as part of the design-time sub-framework depicted in Figure 1.4; (ii) fine tuning of the run-time reliability assessment infrastructure, and (iii) run-time aging assessment, as required by the proposed run-time sub-framework graphically illustrated in Figure 1.5. To this end, we propose to select a minimum-size kernel of critical transistors and based on their status to asses the actual IC degradation and predict its EOL via two methods: (i) as the sum of the critical transistors EOL values, weighted by fixed topology-dependent coefficients, and (ii) by a Markovian framework applied to the critical transistors, which takes into account the joint effects of process, environmental, and temporal variations. The former model exploits the aging dependence on the circuit topology to enable fast run-time reliability assessment with minimum aging sensors requirements. By allowing the performance boundary to vary in time such that both remnant and non remnant variations are encompassed, and imposing a Markovian evolution, the latter probabilistic model can be better fitted to various real conditions. Thus it is mostly meant for design time utilization where it enables: (i) a less conservative guard-band selection allowing the circuit to operate at a higher frequency, when compared to the case when aging is overestimated and thus the circuit performance constrained more than necessary, and/or (ii) the selection of aging mitigation/compensation techniques optimized for the expected working conditions induced degradation. The proposed framework has been validated for different stress conditions, under process variations and aging effects, for the ISCAS-85 c499 circuit, implemented in PTM 45nm CMOS technology. From the total of 1526 transistors, we obtained a kernel of 15 critical transistors, for which the set of topology dependent weights were derived. Our simulation results for 15 critical transistors kernel indicate a small approximation error (i.e., mean smaller than 15% and standard deviation smaller than 6%) for the considered circuit estimated EOL, when compared to the end-of-life values obtained from Cadence simulation, which quantitatively con-

firm the accuracy of the IC lifetime evaluation. Moreover, as the number of critical transistors determines the area and data processing overhead, we also investigated the implications of reducing their number on the reliability assessment accuracy. When reducing the number of sensors by $2/3$, i.e., when only 5 transistors instead of 15 are included into the critical set, the EOL estimation accuracy is diminished with 18%. This indicates that area vs. accuracy trade-offs are possible, while maintaining the aging prediction accuracy within reasonable bounds.

3.1 Introduction

The time dependent drift of critical physical and electrical transistor parameters, due to manufacturing and environmental induced variations, as well as the run-time aging effects, degrades the performance and eventually determines circuit failure.

Most of the past approaches that address the circuit-level reliability analysis mainly focus on either temporal variations - caused by aging mechanism such as Negative Bias Temperature Instability (NBTI), Hot Carriers Injection (HCI), time dependent dielectric breakdown, electromigration, thermal cycling [32, 33], or on process variations [34, 35], without considering the interactions between them. Only recently, studies considering joint effects have been reported in the literature. In the digital domain, aging-aware Statistical Timing Analysis (STA) schemes that rely on analytical expressions of circuit performance features (e.g., propagation delay, signal slope) as a function of process/wearout degradation parameters have been proposed. In [29], based on device parameters statistical spread shifts, the circuit delay fall-out is obtained as an indicator of process variations and NBTI aging effect. In [30] a Statistical Static Timing Analysis method (SSTA) is proposed in order to characterize the circuit delay distribution under process variations and NBTI effects. [28] introduces a statistical age prediction framework for a circuit path under process variations and temporal stress. In [57] an analytical model suitable for circuit level that captures both short term NBTI and process variations effects is developed and used to quantify their impact on the circuit nominal degradation. In [58], the authors introduce the concept of virtual age that reflects the circuit cumulative aging evolution and propose a real time circuit time-to-failure prediction framework.

We note that previous approaches towards aging models are deterministic. However, due to the very nature of the aging inducing phenomena we believe that a more appropriate, but also more complex approach should be a full probabilistic model. In this way the age could be regarded not only as a function of the instantaneous value at time t of a degradation parameter X, for example, but also of its history (from $t = 0$ to the time moment t at which we want to compute the age):

$$A = A(t, x_1, x_2, \dots, x_n), \tag{3.1}$$

where $x_1, x_2, \dots, x_n$ are stochastic processes which enter in the expression of A by their particular realizations. As a consequence, A is also a stochastic process whose

characteristics (e.g., probabilities, moments) have to be obtained from the properties of $x_1, x_2, \dots, x_n$. This is a very general formulation and for a workable model, obviously, we have to impose particular restrictions.

The simplest and roughest simplification of this dependency is to express the age solely as a function of the parameter values at time moment t:

$$A = A(x_1(t), x_2(t), \dots, x_n(t)). \tag{3.2}$$

This brings us back to the point of view adopted in previous deterministic approaches, thus we do not follow this avenue.

Another simplification can be made based on the fact that we don't need all the values between 0 and t but only the values in a finite number of moments. In fact, we can further assume that only the value at the current time moment, (denoted in the sequel by $x_i(t_k)$) and the one at the previous sampling moment (denoted from now on by $x_i(t_{k-1})$) are required. In the general case $x_i(t_k)$ and $x_i(t_{k-1})$ are not independent random variables, but correlated and passing from one to the other could be governed by probabilistic laws. The processes x_i could be Markovian processes and this character could be transferred to A. Moreover, the processes $x_1, x_2, \dots, x_n$ could be correlated. In this case, if we describe (via a change of variables) A as a function of other processes $X_1, X_2, \dots, X_n$ obtained from $x_1, x_2, \dots, x_n$ by a linear transform of Karhunen-Loeve (KL) type [59], the process A can be approximated by making use of a small number of variables. In this manner, one can obtain a correct description of A by, e.g., a function of 4 variables $X_1(k), X_2(k), X_1(k-1), X_2(k-1)$. In view of the above, the following remark is in order: a Markovian model fitted to the age problem must have the transition probabilities not only time dependent but also dependent of the new states. Our approach introduces a Markovian model fitted to the circuit-level aging problem. Furthermore, instead of considering a fixed performance boundary, we allow it to vary in time. In this way we obtain a more flexible model, which takes into consideration that depending on stress duration, the effects on the circuit statistical parameters could be remnant or nonremnant. As a result, guard bands selection and appropriate aging mitigation/compensation techniques, better fitted to real working conditions are enabled.

In view of the previous discussion, this chapter proposes: (i) the selection of a minimum size kernel of critical transistors based on which the circuit end-of-life can be estimated; (ii) a run-time aging framework that estimates the circuit end-of-life as the sum of critical transistors end-of-life values weighted by fixed, topology dependent coefficients; and (iii) a Markovian aging framework that is capable of assessing and predicting the circuit performance degradation and lifetime.

The proposed critical transistors kernel based aging assessment and prediction framework is validated by means of simulation. The simulation is performed in Cadence Relxpert and Spectre, and Synopsys Pathmill, using as test circuit the ISCAS-85 c499, implemented in PTM 45nm technology. Exposing the circuit to several stress profiles, from a total of 1526 transistors, a kernel of 15 critical transistors and their corresponding topology dependent weights were obtained. When subjecting the ISCAS-85

c499 circuit to new sets of stress profiles and comparing the circuit end-of-life estimated with the proposed framework against the results from Cadence and Pathmill, relatively small values of the approximation error (i.e., mean smaller than 10% and standard deviation smaller than 6%) are obtained, which quantitatively validate and confirm the lifetime prediction accuracy of proposed framework. Moreover, as the number of critical transistors determines the area overhead, we also investigated the implications of reducing their number on the reliability assessment accuracy. When only 5 transistors are included into the critical set instead of 15, which results in a 66% area overhead reduction, the EOL estimation accuracy is diminished with 18%. This indicates that area vs. accuracy trade-offs are possible, while maintaining the aging prediction accuracy within reasonable bounds.

The rest of the chapter is organized as follows: Section 3.2 extends the transistor-level aging framework to the circuit level, which is introduced in Section 3.3, and further developed from a probabilistic point of view in Section 3.4, while taking into account the history of aging. The simulation methodology and the obtained results are presented in Section 3.5. The chapter is concluded in Section 3.6 with some final remarks.

3.2 Critical Transistors Kernel Selection

As embedded wear-out sensors are expensive in terms of silicon area and since a circuit may encompass thousands of paths and transistors, a reduction of the number of wear-out measurement sites is thus required for tractability purposes of circuit aging derivation. In the sequel, we present the reduction criteria and the critical paths and critical transistors selection methodologies.

As far as the paths are concerned, we employ as reduction criterion, the path criticality in the circuit from the timing point of view. Specifically, if the aging-induced degradation of a certain path P_1 is larger than that of the initial (unaged, at time 0) critical path P_0 (which determines the clock period), then the circuit timing constraints are violated, and P_1 becomes the circuit new critical path. Therefore, in order to assess the circuit reliability profile, we consider as critical paths the ones that could violate the timing constraints when their comprising transistors are subjected to wear-out induced degradation. By following this principle, the aging of the critical paths can be determined at design-time by performing aging-aware statical timing analysis [60].

As concerns the kernel of critical transistors, we note that for a critical path, only a small percentage of its transistors could potentially cause significant circuit performance degradation due to their aging. As a consequence, a critical path end-of-life can be estimated from a reduced subset of all its comprising transistors, i.e., the path's critical transistors. Thus, the kernel set can be formed as the reunion of the critical transistors for each critical path.

Even though a circuit path may comprise a plethora of transistors, some of them may be weakly correlated with the end-of-life of the critical paths, while others may be

redundant in the estimation if their aging is highly correlated with the aging of other transistors. This suggests the selection of a reduced, common kernel of critical transistors to be utilized for estimating the end-of-life of all the critical paths, as a more appropriate approach. More precisely, we are not interested in selecting the critical transistors that have aged the most, but in selecting the ones that are useful from a prediction point of view, e.g., the redundant but relevant - statistically dependent with the end-of-life of the critical paths - transistors can be excluded from the kernel of critical transistors. In view of the above, we propose to further reduce the cardinality of the critical transistor kernel, by estimating each critical path end-of-life from the same, common subset of critical transistors, regardless of their appurtenance to a particular critical path. That is, instead of using a separate subset of transistors for each path, all of them belonging to the path whose end-of-life is being estimated, we use a common kernel of transistors, not all belonging to the critical path whose end-of-life is being estimated. Thus we select the most relevant critical transistors, specifically the ones with the biggest impact on the circuit aging.

The task of selecting the critical transistor kernel reduces to a multi-response regression problem (i.e., estimating multiple response variables, i.e., paths end-of-lifes, using a reduced, common kernel of input variables, i.e., transistor end-of-lifes). Besides the benefits of critical transistors kernel cardinality reduction, using a single, unified model to estimate all the responses simultaneously exhibits also increased computational efficiency, and better prediction accuracy [61, 62], when compared to building a separate model for each response variable.

The problem of selecting the critical transistors kernel, can be formalized as follows: Suppose we have n end-of-life measurements of the p critical paths and of the m transistors encompassed by the p paths. Let the response variables be denoted by a $n \times p$ matrix $\mathbf{Y} = [\mathbf{y_1} \dots \mathbf{y_p}]$, and the input variables by a $n \times m$ matrix $\mathbf{X} = [\mathbf{x_1} \dots \mathbf{x_m}]$. A linear model of the form:

$$\widehat{\mathbf{Y}} = \mathbf{X} \cdot \mathbf{W} \tag{3.3}$$

is employed for estimating the responses matrix $\mathbf{Y}$, where $\mathbf{W}$ denotes the unknown $m \times p$ regression coefficients matrix desired to have a minimal number q of non-zero rows. Hence q denotes the cardinality of the smallest subset of input variables used to synthesize all response variables. Matrix $\widehat{Y}$ consists of the end-of-life of the critical paths, for the n measurements; matrix $\mathbf{X}$ consists of the end-of-life of the critical transistors, and $\mathbf{W}$ contains the topology dependent weights.

The problem of selecting the kernel of critical transistors and determining the corresponding topology dependent coefficients can be formally stated as follows:

$$\min_{\mathbf{W}} \left\|\mathbf{W}^T\right\|_{l_0} \qquad s.t. \ \frac{1}{2}\left\|\mathbf{Y} - \widehat{\mathbf{Y}}\right\|_F^2 \leq \epsilon, \tag{3.4}$$

where $\|\cdot\|_F$ is the Frobenius norm, that is $\|\mathbf{B}\|_F^2 = \sum_{i,j} b_{ij}^2$, and $\left\|\mathbf{W}^T\right\|_{l_0}$ is the l_0 norm of $\mathbf{W^T}$, defined as the cardinality of the set $\{i \in \{1 \dots p\} : \ w_{i,k} \neq 0 \text{ for some } k\}$.

We note herein that a regression coefficient w_{ij} can be regarded as the importance the i-th input variable has on the j-th response. This optimization problem translates

into minimizing the number of non-zero rows of the regression coefficients matrix $\mathbf{W}$, while keeping the estimation error below a certain bound - in our case, the error tolerance being a function of the circuit timing constraints. Since the norm $\left\|\mathbf{W}^T\right\|_{l_0}$ is a discrete valued function, it yields to a NP-hard problem in terms of computational complexity. The computational intractability can be addressed in two ways: either by using suboptimal algorithms, or by relaxing the problem as for instance via the replacement of the l_0 norm with a convex mixed-norm $l_{p,q}$, defined as:

$$\|\mathbf{B}\|_{l_{p,q}} = \sum_i \|b_{i,.}\|_q^p, \quad \text{where } |b_{i,.}|_q = \left(\sum_j |b_{i,j}|^q\right)^{1/q}, \tag{3.5}$$

among which the most practical instances are $l_{1,q}$ norms with $q \in \{1, 2, \infty\}$ [63–65]. In our case, we use the $l_{1,2}$ norm to quantify the importance of an input variable in synthesizing the response variables. We refer the reader to [66] for the algorithmic details concerning the $l_{1,2}$ optimization problem, and to [67] for the $l_{1,\infty}$ optimization problem.

At this point we have determined the set of critical paths, the set of critical transistors to be monitored by aging sensors, and their topology dependent coefficients w_{ij}. A circuit end-of-life can now be estimated at run-time, as presented in the following section, as the sum of the critical transistors end-of-life values (obtained from the aging sensors), weighted by the fixed topology dependent coefficients w_{ij} determined at design-time with the previously presented methodology.

3.3 Regression-Based Circuit-Level Aging Model

The time-dependent wearout, i.e., aging, affecting a circuit transistors, is reflected at the circuit level as degradation of its performance parameters, such as the increase of the circuit propagation delay. Eventually, the age-induced circuit propagation delay degradation, can exceed the maximum circuit clock period and as a consequence, wrong values may be sampled and hence circuit erroneous functioning induced as the circuit reaches its end-of-life.

In order to estimate a circuit end-of-life, we propose to express: (i) a circuit End-Of-Life (EOL) as the minimum end-of-life of its propagation paths, and (ii) a propagation path end-of-life as the sum of the end-of-life values of all its comprising transistors, weighted by topology-dependent coefficients. Let us consider a circuit and denote by M the number of its propagation paths. The circuit end-of-life can then be expressed as follows:

$$EOL_{circuit} = \min_j \left(EOL_{path\,j}\right), \tag{3.6}$$

$$EOL_{path\,j} = \sum_{i=1}^{N_j} w_{ij} \cdot EOL_i, \tag{3.7}$$

where $j = 1, \ldots, M$, N_j is the number of transistors contained by path j, w_{ij} are topology dependent coefficients, EOL_i represents the end-of-life of transistor i, and $EOL_{path\,j}$ represents the end-of-life of path j.
However, this approach is not feasible, as embedded wear-out sensors are expensive in terms of silicon area and real life circuits may encompass thousands of paths and millions of transistors. A reduction of the number of wear-out measurement sites is thus required for tractability purposes of circuit aging derivation. To this extent, the following model simplifications are made: (i) we reduce the number of paths to a set of critical ones, and (ii) we reduce the numbers of transistors to a kernel set. The model thus becomes:

$$EOL_{circuit} = \min_j \left(EOL_{path\,j}\right) \tag{3.8}$$

$$EOL_{path\,j} = \sum_{i=1}^{N_{reduced}} w_{ij} \cdot EOL_i, \tag{3.9}$$

where $j = 1, \ldots, M_{reduced}$, $M_{reduced}$ is the number of paths, and $N_{reduced}$ is the number of critical transistors. The transistors end-of-life values entering the above equations can be obtained for instance by utilizing the transistor level aging model proposed in Chapter 2.

If an increased accuracy of estimating the circuit end-of-life is required, a probabilistic aging model that takes into account the history of aging is better suited. This is the case of the Markovian model presented in the next section, which can estimate the circuit end-of-life based on the kernel of critical transistors obtained according to the methodology introduced in Section 3.2.

3.4 Markovian Circuit-Level Aging Model

In the proposed framework, we define the age of a circuit as a function of many parameters which can be divided into three main categories: (i) **d**, design parameters (e.g., channel width W), which are subject to optimizations; (ii) **s**, statistical parameters (e.g., threshold voltage V_{th}) that fluctuate during to manufacturing process but also evolve in time depending on the dynamic operating conditions - their random behavior can only be described in probabilistic terms as random processes; and (iii) **r**, range parameters (e.g., temperature T, supply voltage V_{DD}) whose variations are handled by specifying the range of values that can be attained.

In the following, the relation between the degradation parameters X_i and the performance parameter P_{out} will be given by a function f:

$$f : \mathbb{R}^n \to \mathbb{R};\ f(\mathbf{X}) = P_{out}.$$

During the lifetime of a device, its performance has to be better than an imposed value, which in our case means:

$$P_{out} > P_{out\,min}. \tag{3.10}$$

As P_{out} is time dependent (more precisely decreases with increasing time) through various parameters among which, some are random processes, the lifetime of the device can be expressed in probabilistic terms as:

$$R(t) = Prob\{P_{out}(t) > P_{out\,min}\}, \tag{3.11}$$

The device end of life is thus given by the value of t for which $P_{out}(t) = P_{out\,min}$.

Further, we adopt the usual method to achieve tractability of our problem, namely, in the case of more than one scalar statistical parameter, we apply on each of these parameters (with the restriction of having unimodal distributions) appropriate transforms to convert them into normal distributed random variables [68], while maintaining the correlation among each pair. In consequence, the statistical parameters become a normal distributed vector. In this way, we are able to compute the worst-case distance d_w, defined in [69] as the Mahalanobis distance between the mean point $\mathbf{s}_0$ and the worst case point $\mathbf{s}_w$ (i.e., the point belonging to the set of all parameters that violate a specification and is closest to the mean vector $\mathbf{s}_0$):

$$d_w^2 = (\mathbf{s}_w - \mathbf{s}_0)^T \cdot \mathbf{C}^{-1} \cdot (\mathbf{s}_w - \mathbf{s}_0), \tag{3.12}$$

where, as said, $\mathbf{s}$ denote the vector of statistical parameters after transforming them into Gaussian variables; $\mathbf{s_0}$ is its mean vector, and $\mathbf{C}$ is its covariance matrix, all at the same time moment. The worst case distance d_w, obtained with Equation (3.12) (based on the fact that the level contours are ellipsoids) is a measure of the circuit robustness. The worst case point $\mathbf{s}_w$ is found by solving:

$$\mathbf{s}_w = \text{argmin}\,(d_w^2 \mid P_{out} = P_{out\,min}).$$

As time t increases, the probabilistic properties of the statistical parameters vector, $\mathbf{s}(t)$, evolve, i.e., the mean $\mathbf{s}_0$ and the covariance matrix $\mathbf{C}$ are functions of time and the worst case distance becomes smaller. In spite of the fact that $\mathbf{s}_0$, the mean, is inside the admissible region, characterized by $P_{out} > P_{out\,min}$, the reliability with the new worst case distance attains its minimum acceptable value and the circuit reaches its end of life. This evolution is graphically caught in Figure 3.1.

Actually, the performance boundary defined as $P_{out}(t) = P_{out\,min}$ in the space of $\mathbf{s}$ coordinates, and being characterized by the performance function $P_{out}(\mathbf{d}, \mathbf{s}(t), \mathbf{r})$ which depends on t only via $\mathbf{s}(t)$, could be a too restrictive model for what one might encounter in real situations. For instance, if the range parameters vary in time too, this variation could have remnant - or only transient if the circuit was not exposed for a long time - influence on the physical modifications of the devices. This situation is easier described by allowing the performance boundary to vary in time, as graphically illustrated in Figure 3.2. Therefore, we propose to employ a space with one more coordinate - time - and represent the evolution of the reliability ellipsoids as a tube and the evolution of the performance boundary as a surface exterior to the tube. In this way, both situations are encompassed, that is when the device degradations with T and V_{DD} variations for instance, are remnant, and when they are not remnant.

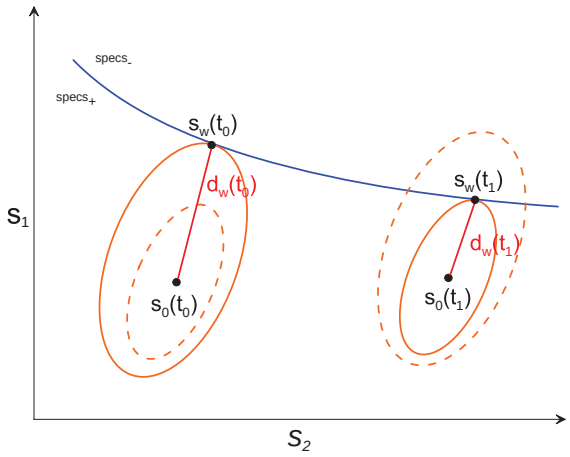


Figure 3.1: Lifetime Evolution for Fixed Performance Boundary Graphical Representations.

At this stage we have a model in which the performance scalar P_{out} depends on the vector of statistical parameters **s**, which has normal and correlated components. Therefore, an orthogonal transform (e.g., KL [59]) can be applied to decorrelate them; after this step the next simplification is to maintain only the two most important components and neglect all the others. It should be noted that these two most important components, retained, are now uncorrelated and as a consequence independent. The following step in developing a workable model is to accept a Markovian evolution and obtain the new values of the two components by applying the transition matrix to the old ones. In fact we deal with two uncorrelated Markov chains, with each component evolving separately. Both processes have a continuous space of states, $\mathbb{R}$, the set of real numbers. In consequence, the probability of a value is obtained by the Chapman-Kolmogorov equation as an integral over $\mathbb{R}$ from the conditional probabilities of that value, given each of the possible previous values:

$$p_{k+1}(y) = \int_{x \in \mathbb{R}} p_k(x) \cdot p_k(y\,|x)\,\mathrm{d}x.$$

In the simplest case, i.e., a stationary Markov process, the model assumes a transition probability that is time independent, that is to say $p_k(y\,|x) = p(y|\,x)$. As previously stated, the evolution of s_1 is independent of s_2. In computing the evolution of the Probability Density Function (PDF) of $s_1(t)$ and $s_2(t)$, we shall replace the continuous time t with a discrete set of integers k. The PDF of $s_1(k+1)$ can be obtained from the PDF of $s_1(k)$ (the same reasoning holds true for s_2) by an integral formula where the Markovian character has to be defined so as to fit the simulation results. This approach is more general than the one developed in [70] and thus can be fitted better to various real conditions.

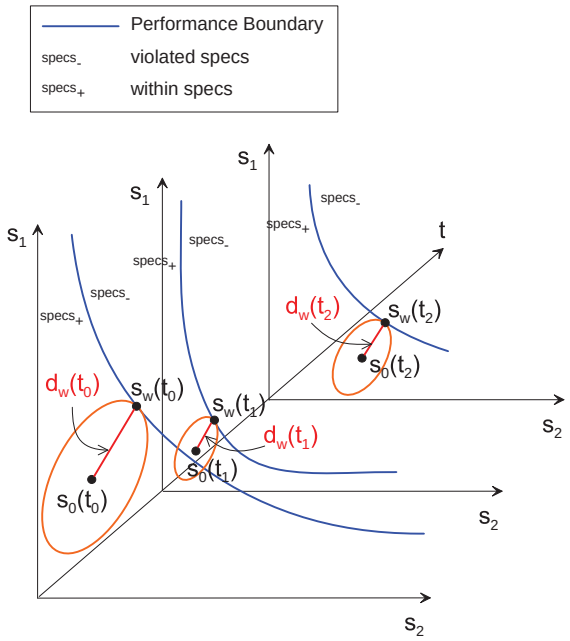


Figure 3.2: Lifetime Evolution for Time Varying Performance Boundary Graphical Representations.

The two independent Gaussian processes, s_1 and s_2, to which we impose a Markovian character, are therefore Wiener processes. The time evolution of their PDF-s for continuous time is described by [71]:

$$p(s_{i,0}, s_i; t)\, ds_i = Prob\{s_i < s_i(t) \leq s_i + ds_i \mid s_i(0) = s_{i,0}\}$$
$$= \frac{1}{\sigma_i\sqrt{2\pi t}} \cdot \exp\left\{-\frac{(s_i - s_{i,0} - \mu_i t)^2}{2\sigma_i^2 t}\right\} ds_i,$$

where $i \in \{1, 2\}$ and μ_i and σ_i denote the mean and the variance, respectively, of the two processes.

The boundary of the permissible domain in the $(s_1,\ s_2)$ plane is known and given by the functional relation between P_{out} and the two statistical parameters s_1 and s_2 (see Chapter 2). As s_1 and s_2 are independent processes, their bi-dimensional PDF is the product of their one dimensional PDF-s. Along the time, the mean (the drift) and variance evolve as for the Wiener process and specifically, increase proportional with t. The circuit starts its life with given values $s_{1,0}$ and $s_{2,0}$ in the admissible domain; as t increases, the mean as well as the variance increase and the point $(s_1,\ s_2)$ eventually reaches the border. Actually we cannot wait until this event happens: we have to fix the moment when the probability $Prob\{P_{out}(s_1,\ s_2) < P_{min}\}$ and this probability is given by the probability that $(s_1,\ s_2)$ is out of the border.

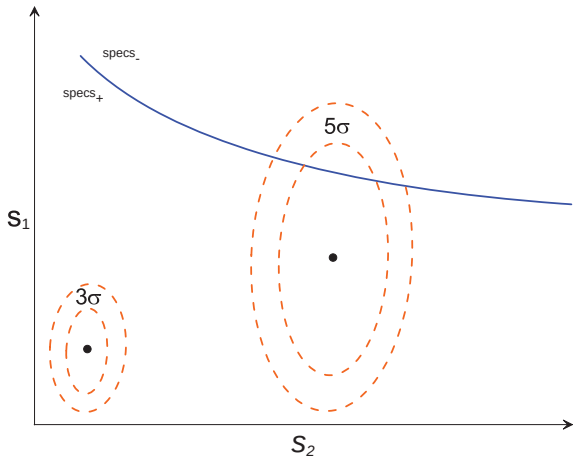


Figure 3.3: 2D-PDF Evolution.

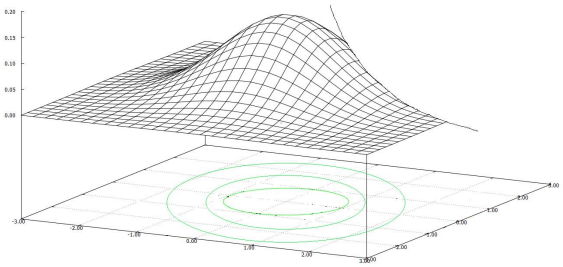


Figure 3.4: 2D-PDF Truncated by Performance Boundary.

In Figure 3.3 is presented a sketch of this situation in two successive moments with the 2-dim PDF of $(s_1,\ s_2)$, while in Figure 3.4 is depicted the 2-dim PDF cut away by the performance specifications. Given the Gaussian character of our variables one can compute the probability of the domain D_{ext} (out of the performance border) for any time. For an estimation of the moment we are interested in, it is enough to have the values of:

$$Q(t) = Prob\{(s_1(t),\ s_2(t)) \in D_{ext}\}$$

in a finite number of moments and use a linear interpolation between them. In this way (the intersection of $Q(t)$ with the horizontal line $Q = Q_{max}$ admissible) we obtain the moment when the circuit reaches its EOL.

We stress out that using the hypotheses mentioned above we are able to bypass the difficulties of a direct Markovian model (in our model the Markovian character is included in the Wiener model for which there are classical results). The parameters of the Wiener processes have to be obtained from simulations.

When the $\mathbf{r}$ parameters are varying too, it is necessary to move the border accordingly at the same time as the PDF of $(s_1,\ s_2)$ is evolving in the $(s_1,\ s_2)$ space. There are two situations: either one knows their variation or only a PDF of this border (and so of the domain D_{ext}), is known. In the last case we have to compute $Q(t)$ for any position of the border - we shall index the possible positions at time t by a variable u - and to obtain the probability we look for as a weighted value of the probabilities for each D_{ext}:

$$Q(t) = \int Q(t,\ u) \cdot Prob\left\{D_{ext}(u)\right\} \mathrm{d}u.$$

It is very likely that we have only a few values of the $\mathbf{r}$ parameters, as for instance three values of V_{DD} with probabilities p_1, p_2, and p_3; for a time t. In such a case $Q(t)$ can be obtained as:

$$Q(t) = p_1 \cdot Q(t,\ D_{ext1}) + p_2 \cdot Q(t,\ D_{ext2}) + p_3 \cdot Q(t,\ D_{ext3}),$$

where $p_1 + p_2 + p_3 = 1$. We note inhere that this formulation does not contain the case when a variation of $\mathbf{r}$ parameters induces modifications on other parameters of the function P_{out}. In such a case, the Wiener process model parameters have to be continuously adapted during run-time.

3.5 Performance Evaluation

In this section, the framework with fixed topology-dependent weights and the Markovian circuit-level aging framework are validated and their end-of-life estimation accuracies are evaluated. The simulation is conducted on the ISCAS-85 c499 circuit, which is a single-error-correcting circuit with 41 inputs, 32 outputs, and 202 gates, using PTM 45nm technology. The reliability analysis (BTI and HCI aging) is carried in Cadence RelXpert and Virtuoso Spectre simulators [55], using the AgeMOS model extracted in BSIMPro+ [72] for PTM 45nm technology [73]. The transistor-level static timing analysis in performed in Synopsys Pathmill [74].

The validity of estimating a circuit end-of-life from the end-of-life of the critical transistors in the kernel set, is examined by exposing the circuit to several stress profiles (e.g., varying duty-cycle, temperature, input vectors). Based on each profile's fresh and aged timing reports, we determine the set of aging critical paths, i.e., we select the paths with propagation delay exceeding the clock period. In our case we impose an end-of-life target of 10% propagation delay degradation, and retain the first 100 critical paths. The initial set of transistors that constitute the 100 critical paths and which is to be reduced to a set of critical ones, consists of 53 transistors. Then, according to the methodology described in Section 3.2, the regression matrix is derived, and implicitly the reduced set of critical transistors. Figure 3.5 illustrates the regression matrix obtained for the analyzed circuit. The input and output variables are the end-of-life of the critical transistors and the end-of-life of the critical paths, respectively, obtained from simulation. Based on the input and output variables, the regression

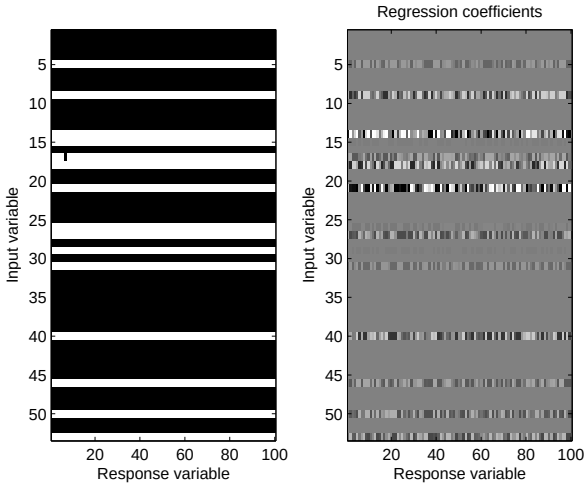


Figure 3.5: The Regression Coefficients Determining the Reduced Set of Critical Transistors.

coefficients, i.e., the topology dependent weights, are obtained using the model from Section 3.2. In the left subfigure, the input variables that are discarded from the model are represented in black, while the reduced set of inputs - in our case 15 from a total of 53 - that are relevant for synthesizing the output responses - in our case 100 aging critical paths - are represented in white. The right subfigure depicts in grayscale the variable regression coefficients w_{ij} corresponding to each relevant input variable, for all the output responses.

Having determined the minimum-size kernel of critical transistors and their topology dependent coefficients, we are now in the position to validate the resulted model for a new set of input aggression profiles, using Equation (3.3). Figure 3.6 illustrates the normalized simulated circuit end-of-life values vs. the normalized estimated circuit end-of-life values in the case of the new set of input aggression profiles. The simulation results reveal a mean estimation error of 15% and a variance of 6%, which confirms that the determined kernel of critical transistors can be utilized to estimate the circuit end-of-life at run-time fairly accurate. A remark is in order: To achieve a good estimation of a circuit end-of-life, besides the matter of choice of solving the regression problem, the initial sampling for multiple levels of stress should be carefully considered.

Since the reliability aware management of integrated circuits implemented in advanced technology nodes requires reasonably accurate but fast run-time reliability profiling, a further reduction of the number of aging measurement sites could be desired. To this extent, we study the trade-offs between the number of critical transistors that are used for end-of-life circuit estimation, and the circuit end-of-life estimation accuracy. Figure 3.7 depicts the error analysis of the circuit end-of-life, for different

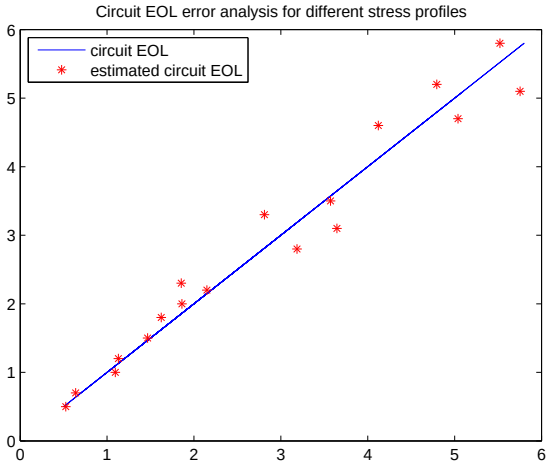


Figure 3.6: Error Analysis of Circuit End-Of-Life Estimation Based on the End-Of-Life Values of the Critical Transistors.

subsets - with different cardinality - of critical transistors, when subjecting the circuit to 5 new stress profiles. For each stress profile, 5 subsets of critical transistors with different cardinalities, which are obtained by reducing the initial critical transistors kernel with 2%, 5%, 10%, 25%, and 30%, are being considered. The percentage of estimation accuracy loss is reported relative to the estimation accuracy obtained when using the entire kernel of critical transistors. The transistors are eliminated based on their relevance in estimating the circuit end-of-life (i.e., the less relevant goes out first). We observe a similar trend of the end-of-life circuit estimation quality loss when decreasing the number of critical transistors for all considered stress profiles. As concerns the differences in the rate of estimation accuracy loss, they can be attributed to the relevance of the dropped transistors in estimating the model responses for considered input stress profiles. However, taking into consideration that in most situations a very precise estimation of the circuit end-of-life is not required, a coarse reliability assessment is sufficient to enable graceful performance degradation and prolong the circuit lifetime via aging mitigation and compensation techniques. One can observe in Figure 3.7 that for the considered circuit, reducing the number of sensors by $2/3$ (5 sensors instead of 15 to monitor the reliability of a 202 gates circuit) can be achieved for less than 18% loss in circuit end-of-life estimation accuracy (reported relative to the estimation accuracy achieved by employing the entire kernel of critical transistors), which makes it a potentially feasible approach for practical implementations.

For a more accurate estimation of the circuit end-of-life, that takes into account the history of aging, we apply the Markovian framework on the kernel of critical transistors previously validated. For the purpose of illustration, we employ Monte Carlo simulation loops, approach which is typical for analog circuits, where the analytical

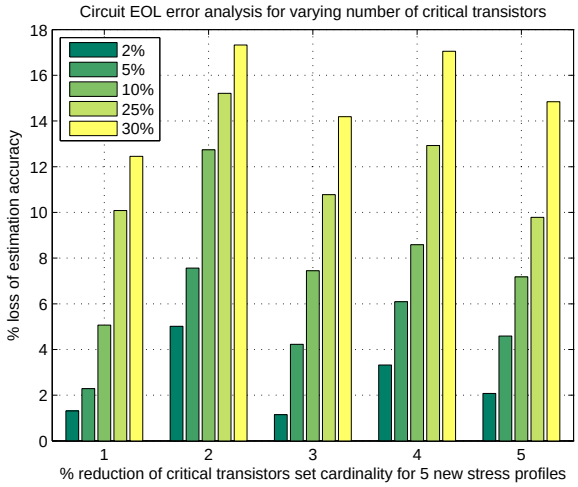


Figure 3.7: Error Analysis of Circuit End-Of-Life Estimation Based on the End-Of-Life Values of the Critical Transistors.

expressions of circuit performance features as functions of statistical parameters are not known. We choose as circuit performance metric the propagation delay. As concerns the statistical parameters, we use the threshold voltage, V_{th}, the low-field mobility μ_0, the oxide thickness t_{ox}, and the oxide capacitance C_{ox}. After decorrelation, the components V_{th} and μ_0 are retained.

In Figure 3.8, is depicted the normalized circuit end-of-life, which is defined as the time when the propagation delay is degraded by $v\%$. For expository purposes we define the end-of-life target for the considered simulation framework as $v = 10\%$ degraded propagation delay. We consider several stress profiles (e.g., varying duty-cycle, temperature, input vectors), and obtain the corresponding performance boundary for defined end-of-life target in the $(\mathbf{V_{th}}, \mu_0)$ space, as result of reliability analysis (NBTI and HCI aging) and Monte Carlo simulation. For each profile and corresponding data set of statistical parameters, we determine $Q(t)$ in a finite number of moments, interpolate them and estimate the end-of-life time moment. This is compared against the accurate end-of-life value which is obtained by means of simulation, i.e., the time moment when $\mathbf{s} = \mathbf{s_w}$, for the obtained performance boundary. Figure 3.8 illustrates the obtained circuit end-of-life prediction accuracy using the Markovian framework on the set of critical transistors. We obtained an approximation error with mean $(< 10\%)$ and standard deviation $(< 15\%)$. As expected, the estimated end-of-life values are further from the values obtained with Cadence. We attribute this to the Markovian approach and to the fact that we use multiple monitors to quantify the aging process. In fact, as the Markovian model takes into consideration more parameters and aging sources, these estimated end-of-life values may be closer to the real end-of-life values but for the time being we do not have the means to validate this conjecture. The

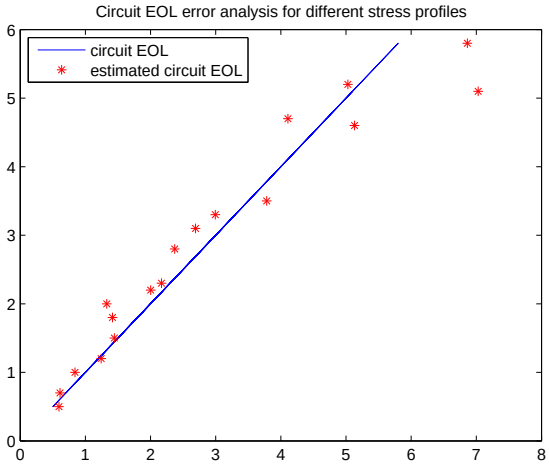


Figure 3.8: Error Analysis of Circuit End-Of-Life Estimation, Using the Markovian Statistical Framework on the Set of Critical Aging Transistors.

proposed Markovian framework necessitates the monitoring of multiple degradation parameters per transistor, e.g., V_{th}, μ, and hence multiple sensors are required for one transistor. This makes this approach less feasible for run-time aging assessment and prediction, and better suited at design-time, enabling a robust, fast and accurate aging evaluation, which takes into account the history of the degradation caused by joint effects of process, environmental, and aging-induced variations.

Furthermore, the proposed Markovian framework is general, and hence suitable for emergent nanoscale technologies, under the provision that the technology is known (and implicitly the afferent design and statistical parameters are known). In this work, we particularized the framework for the bulk CMOS in 45nm technology node, using certain commonly employed parameters such as V_{th} as statistical parameter. However the same line of reasoning can be applied for newer technologies to similar or different electrical and/or performance parameters. As far as the topology dependent framework is concerned, it can also be applied for emergent nanotechnologies, even if the circuit critical paths may have different constituent blocks instead of the bulk CMOS transistors that we employed for expository purposes.

3.6 Conclusion

In this chapter we introduced a model able to infer circuit aging based on the aging of comprising transistors. Specifically, we proposed to predict the circuit End-Of-Life (EOL) as a sum of the critical transistors individual EOL values, weighted by circuit topology dependent coefficients that are fixed at design-time via a multi-response re-

gression approach. Given that abstracting the circuit level aging status from the aging of all its component transistors is not an option for non-toy circuits we further propose a methodology to select a reduced set of circuit transistors, i.e., those which are critical from the aging point of view, based on whose degradation circuit aging assessment becomes computationally tractable. To this end, we identified the circuit paths which are critical from the aging point of view, i.e., the paths whose aging induced degraded performance violates their un-aged timing requirements. In view of the fact that from each such aging critical path, only a small percentage of transistors could potentially cause the greatest extent of damage we introduce a method to identify and select those as constituents of an aging critical transistors kernel. The transistors thus selected are the ones to have their aging monitored, and used to infer the aging of the overall circuit via the proposed topology dependent circuit-level model. When applied on the ISCAS-85 $c499$ circuit implemented in PTM 45nm CMOS technology our approach identified a 15 critical transistors kernel out of a total of 1526 transistors, which greatly benefits from the standpoints of wear-out measurements sites (the number of aging sensors) reduction and circuit aging derivation tractability. When inferring the circuit age with the proposed model based on the 15 identified critical transistors, we obtain a small circuit aging approximation error (mean $< 15\%$ and standard deviation $< 6\%$) w.r.t. the reference circuit aging obtained by means of Cadence simulation, which quantitatively confirms the evaluation accuracy of our approach. When reducing the number of sensors by $2/3$, i.e., only 5 transistors out of the kernel of 15 are utilized for aging evaluation, the EOL estimation accuracy is diminished with 18% (reported relative to the estimation accuracy achieved by employing the entire kernel of 15 critical transistors), which indicate that trade-offs between area overhead and EOL estimation accuracy are possible, the lifetime prediction being kept within reasonable bounds. The proposed circuit aging assessment approach is deterministic and fast, and requires a minimum number of in-field aging sensors (to monitor the circuit kernel of critical transistors), which makes it a good candidate for run-time aging evaluation.

We further proposed a fully probabilistic circuit level aging model which regards age not only as a function of the instantaneous degradation, but also of past degradation. We introduced a model that allows the circuit performance (e.g., clock frequency) to vary in time, and has a Markovian character (the age being expressed as a function of the degradation parameters values at the current sampling moment and the ones from the previous sampling moments), as opposed to the traditional deterministic approach, which considers a fixed performance boundary and includes only the current sampling moment degradation values when inferring the age. In this way, we account for both remnant and non-remnant degradation effects depending on the circuit stress duration. To deal with aging, a guard-band is typically selected at design time (i.e., the circuit maximum frequency is lowered to account for the worst case of performance loss due to aging). When employing our aging assessment method, the aging is estimated for expected typical circuit working conditions, allowing thus (i) a less conservative guard-band selection (a higher allowable circuit maximum operating frequency), and (ii) aging mitigation/compensation techniques optimized for the ex-

pected working conditions induced degradation. Simulation results performed for the ISCAS-85 $c499$ circuit implemented in PTM 45nm CMOS technology (with a kernel of 15 critical transistors out of 1526 transistors) yielded an approximation error with mean $< 10\%$ and standard deviation $< 15\%$ w.r.t. the counterpart values provided by Cadence simulations. As expected, the aging values obtained with our framework are further from the Cadence values, as we take into account more parameters and aging sources; in fact we are probably closer to the real EOL values (but we have no means to verify this conjecture). We note that the proposed Markovian aging model necessitates the monitoring of multiple degradation parameters per transistor, e.g., V_{th}, μ, and hence multiple sensors are required for one transistor, which makes this approach less feasible for run-time aging assessment and prediction, but better suited at design-time, enabling a robust, fast, and accurate aging evaluation.

Having discussed thus far a framework to assess the reliability status of a circuit based on the reliability figures of its comprising devices, we shall present in the next chapter an aging sensor design suitable for monitoring the transistors/circuit blocks degradation.

Note. The content of this chapter is based on the following papers:

N. Cucu Laurenciu, S. D. Cotofana, **A Markovian, Variation-Aware Circuit-Level Aging Model**, Proceedings of the 8^{th} IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH 2012), pp. 116–122 (2012).

N. Cucu Laurenciu, S. D. Cotofana, **Critical Transistors Nexus Based Circuit-Level Aging Assessment and Prediction**, Journal of Parallel and Distributed Computing 74(6), pp. 2512–2520 (2014).

4

Aging Sensor Design for Dynamic Lifetime Resources Management

Accurate and efficient degradation measurements are highly desirable as underlying key information in order to enable an effective IC lifetime management. Though, up to date, various aging sensors have been proposed only few of them can provide real quantitative aging measurements. In view of this, we propose a direct measuring scheme, which makes use of the drain current as aging indicator. We design a novel on-chip aging sensor able to capture the amalgamated degradation effect of various failure mechanisms on the IC under observation aging status and reliability. We achieve this by detecting the peak power supply current (I_{pp}) degradation (which is a signature of the total drain current) and converting it into a pulse width modulated signal, further used to extract the overall circuit aging status. Unlike existing aging sensors, which estimate the aging status indirectly by replicating devices and exposing them to identical stress as the computationally active ones, the proposed sensor allows for direct aging assessment for a single device and/or large circuit blocks. We note that unlike existing sensors, which monitor the aging of individual transistors, the proposed sensor can monitor in principle the entire circuit, which means that no transistor kernel identification and models for aging abstractization from transistor to circuit level are any longer required. Accelerated Cadence test simulation for a set of ISCAS-85 benchmark circuits using TSMC 65nm CMOS technology, indicates that the drain current exhibits a similar aging rate as the threshold voltage for the most of circuit lifetime. However, towards circuit EOL, the peak power supply current and the threshold voltage aging indications start to deviate from each other. Unfortunately, no other aging reference is available thus we cannot tell which method provides better estimates for that circuit lifetime part. We note however that high accuracy is essential during an IC early life, when it undergoes the greatest wear-out induced degradation, demanding thus accurate and fast reliability assessment and EOL prediction for an appropriate degradation compensation/mitigation measure.

4.1 Introduction

Usually, Integrated Circuits (ICs) lifetime requirements are mostly made based on worst-case assumptions, which leads to highly conservative margins on technology parameters, resulting in the under utilization of the technology potential. To make better use of the technological improvement the pessimistic assumption should be relaxed and combined with a Dynamic Reliability Management (DRM) framework that relies on online sensors to measure the ICs aging status. In the recent past, a number of approaches for aging/reliability monitoring have been reported. In [39], Kim et al. introduced an on-chip aging monitor for high resolution degradation measurements by detecting beat-frequency from a pair of ring oscillators. Keane et al. further extended this idea to an "all-in-one" sensor for BTI, HCI, and TDDB in [40]. Though high precision can be achieved by their circuitry, a large area overhead is required (0.035mm^2 in 130nm technology). Karl et al. proposed compact in-situ sensors for monitoring NBTI and TDDB, respectively, in [41]. These sensors work in the sub-threshold region with leakage current to increase the sensitivity. Even though they require a small area overhead these sensors are sensitive to process, voltage, and temperature variations. Agarwal et al. proposed aging sensor designs integrated inside a flip-flop to detect delay violation(s) in [42, 43]. These designs are relatively small and can be potentially included in many chip flip-flops. However, this kind of sensor can only check delay violation in a static or quasi-static time window ("guard band"), and thus no quantitative aging information can be collected.

Nevertheless, all these sensors above have a common shortage that they cannot provide a direct measurement of the real aging status of the Circuit Under Observation (CUO). Previous work can be divided into two groups: (1) sensors that use performance comparison of fresh and stressed devices to get aging information; and (2) sensors that use timing violation checking in a predefined "guard band". For the former group, the aging information is extracted from an additional stressed device, which is carefully placed to make it exposed to the same stressing environment as the CUO. Though high correlation can be achieved by a smart enough placement algorithm, such an approach increases the complexity and effort at design-time and still ends up with an indirect aging measurement. The latter group of sensors can detect the real aging of the CUO, however, they cannot give a quantitative measurement on aging, as mentioned in the above paragraph.

To overcome the common shortage of the existing sensors, we propose a novel online aging sensor able to directly measure the real circuit degradation under NBTI and HCI. The proposed sensor measures the CUO I_{pp} value and converts it into a Pulse-Width Modulated (PWM) signal. The I_{pp} value accurately reflects the aging information as its value is affected by the degradation of multiple aging sensitive device parameters such as the threshold voltage (V_{th}) and the carrier mobility (μ).

The major contributions of this chapter can be summarized as follows:

- We derive the relationship between peak power supply current I_{pp} degradation and multiple failure mechanisms and demonstrate that it can accurately capture

the effect of multiple aging mechanisms;

- We propose a novel aging sensor, which is capable of directly measuring the real IC aging status caused by the amalgamated effect of NBTI and HCI failure mechanisms;
- With a V_{th} sensor, only one transistor can be observed when the entire circuit is monitored with one sensor. The proposed method allows us to observe the entire circuit instead of a transistor, which substantially reduces the area overhead, alleviates the problem of finding the optimum location for the sensors and to extrapolate the overall circuit-level aging from the transistor-level aging;
- We evaluate our proposal by means of simulations. We utilized a set of ISCAS-85 benchmark circuits to validate our choice of using the drain current to monitor their aging status. Simulation results indicate similar aging trends for both the drain current I_D and the threshold voltage V_{th}, but with the I_D exhibiting better sensitivity as the circuits approach their EOL.

The rest of the chapter is organized as follows: Section 4.2 presents the aging model describing the relationship between the peak current value and the IC aging conditions under multiple failure mechanisms. In Section 4.3, the direct aging measurement framework is presented, followed by the description of the I_{pp}-based sensor architecture and the corresponding performance evaluation. Section 4.4 validates experimentally the aging estimation accuracy and the feasibility of our proposal - using the drain current for direct aging monitoring. Section 4.5 concludes the chapter with a summary of this work.

4.2 Power Supply Current as Aging Monitor

In prior work, the threshold voltage (V_{th}) is the most common physical parameter selected as indicator of transistor aging progress. However, aging sensors based on V_{th} can only be used to monitor the aging status at transistor level. Thus, for circuit level aging measurements many such sensors are required. Moreover, sensor positioning and the extrapolation method that can bring aging information from transistor to circuit level are far from being trivial issues. Furthermore, V_{th} is hard to be extracted directly without interrupting circuits' normal function. As a result, all V_{th} based aging sensors make use of sacrificed devices in order to replicate the stress to which the circuit under observation is subjected to, leading to an indirect measurement. In order to measure the real aging status inside a circuit, we propose to use the power supply current (I_p) as aging indicator. Power supply current I_p is the total drain current passing through the supply voltage terminals. Without loss of generality, we start to determine the relationship between I_p degradation and failure mechanisms on an inverter circuit. The operating regions of each transistor during a low to high input transition (i.e., the input voltage switching from "0" to "1"), are graphically illustrated in Figure 4.1(d). During this progress, at some point ($V_{in} = V_{out} = V_M$), both transistors are saturated and the power supply current I_p reaches the maximum value I_{pp} because the V_{DS}

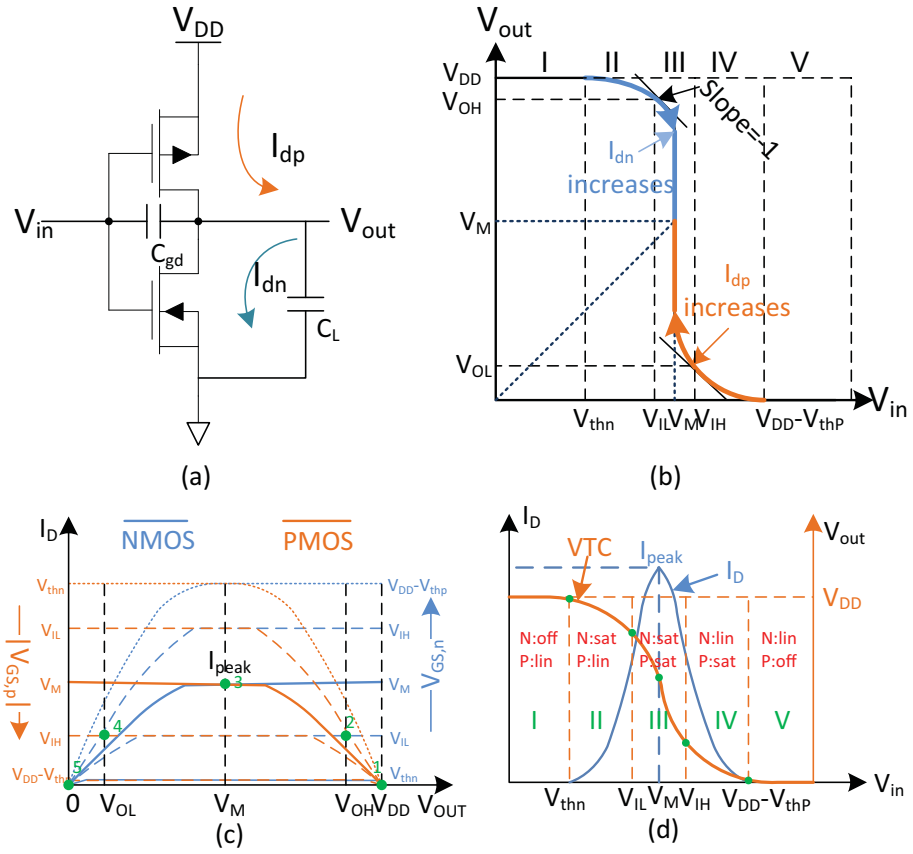


Figure 4.1: Inverter Peak Current: (a) The Circuit Model; (b) VTC and the Operating Regions of PMOS, NMOS Transistors; (c) Intersections of the Output Characteristic Curves of PMOS, NMOS Transistors; (d) The Peak Power Supply Current and Transistor Operating Regions.

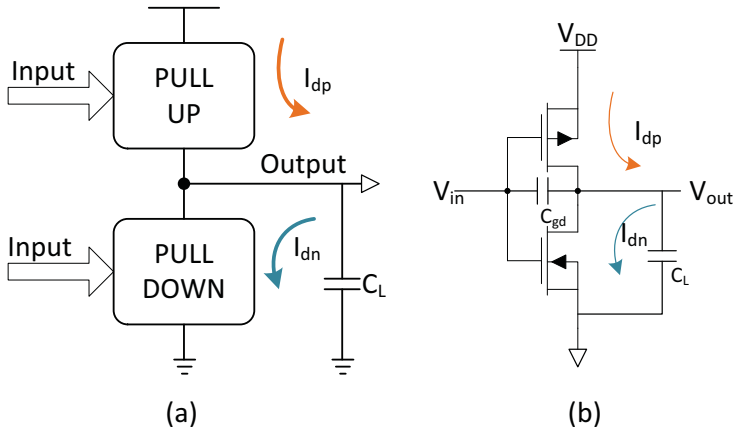


Figure 4.2: Peak Current of CMOS Logic: (a) A General Illustration of CMOS Network; (b) Equivalent Invert Circuit for the Pull-Up Network.

values are equal for both transistors. Further increase of input voltage will make the NMOS enter the linear region and as a consequence the drain current I_D decrease until 0 (because the PMOS is switched-off by the gate overdrive voltage). The analysis is similar for the high to low input transition (when the input voltage changes from "1" to "0"). Summarizing, the current I_p reaches a peak I_{pp} during the input signal transition, when $V_{in} = V_{out}$.

The above analysis can be easily applied to more complex CMOS networks such as the general CMOS logic structure, depicted in Figure 4.2(a). The PMOS devices are equivalent to a pull-up network and the NMOS devices are equivalent to a pull-down network. Since the input vectors to the pull-up and the pull-down networks are complementary in CMOS logic, the working regions of pull-up and pull-downs networks shift oppositely during input signal transitions. Thus at some point, the I_p of the entire network will reach a peak I_{pp}. For the fresh (unaged) devices, this peak value is constant for a given input pattern, so the degradation of the peak current can be chosen as indicator for assessing the aging status inside any (large) CMOS logic network.

Based on the above analysis and since the global I_{pp} is just a special case of I_D (i.e., when the total drain current from both the pull-up and the pull-down networks reaches its maximum), we can consider the peak power supply current as a signature of the drain current. Therefore, without loss of generality, we shall derive subsequently the aging model for I_{pp} using the drain current.

Generally, the drain current in the saturation region can be expressed as:

$$I_{D_{sat}} = \mu C_{ox} \frac{W}{L} (V_{GS} - V_{th})^2. \tag{4.1}$$

A more general form of the drain current can be expressed as:

$$I_D = \mu C_{ox} \frac{W}{L} \left[(V_{GS} - V_{th}) V_{DS} - \frac{V_{DS}^2}{2} \right]. \tag{4.2}$$

(4.1) and (4.2) indicate the relationship between I_D and key device parameters, such as the mobility μ, the device threshold voltage V_{th}, and the oxide capacitance C_{ox}, which are degrading under joint NBTI and HCI induced stress. The peak power supply current is a special case of I_D, i.e., the I_D value which is equal in both the pull-up and the pull-down network. Therefore, as far as aging effects caused by amalgamated wearout mechanisms are concerned, I_{pp} follows the same rule as I_D does.

NBTI is an intrinsic front-end-of-line wearout mechanism which occurs in PMOS transistors mainly when the gate is subjected to a negative input voltage. The NBTI-induced I_D damage includes trap generation at the channel-dielectric interface and as well as inside the bulk of the dielectric. Consequently, the threshold voltage V_{th} shifts and channel mobility μ degrades due to the trap generation. The V_{th} degradation induced by NBTI can be expressed by:

$$\Delta V_{th} = \frac{q \Delta N}{C_{ox}}, \tag{4.3}$$

where q is the electron charge and ΔN is the total trap density generated by NBTI. The mobility degradation can be described by the following equation [75]:

$$\mu = \frac{\mu_0}{1 + \alpha \Delta N}, \tag{4.4}$$

where μ_0 is the original channel mobility, α is a process-dependent constant, and $\alpha \approx 2.4 \times 10^{-12} cm^2$ [76]. For small ΔN, using $\Delta \mu = \mu_0 - \mu$, the channel mobility degradation can then be estimated by:

$$\frac{\Delta \mu}{\mu} = \frac{\Delta N}{1 + \alpha \Delta N} \approx \alpha \Delta N. \tag{4.5}$$

Assuming the threshold voltage shift and channel mobility degradation are independent progresses, then the change of I_D can be expressed as:

$$\Delta I_D = \frac{\partial I_D}{\partial V_{th}} \Delta V_{th} + \frac{\partial I_D}{\partial \mu} \Delta \mu. \tag{4.6}$$

Since the NBTI stress happens when PMOS is in saturation mode, applying the above equation to (4.1), yields:

$$\Delta I_D = \frac{I_{D0}}{(V_{GS} - Vth - V_{DS}/2)} \Delta V_{th} + \frac{I_{D0}}{\mu_0} \Delta \mu. \tag{4.7}$$

Generally, $V_{DS} \approx 0$ when the channel is conducting. Inserting (4.3) and (4.5) into above equation, the degradation of I_D due to NBTI can be estimated as follows:

$$\frac{\Delta I_D}{I_{D0}} = \frac{q \Delta N}{C_{ox}(V_{GS} - Vth)} + \alpha \Delta N, \tag{4.8}$$

which suggests that the fraction of I_D degradation is proportional to the severity of NBTI wearout.

Similarly, the HCI-induced degradation is also a combined effect of threshold voltage shift and channel mobility degradation, which implies that basically (4.6) also holds true for HCI degradation. Hence, the amalgamated aging effect of NBTI and HCI can be written as:

$$\frac{\Delta I_D}{I_{D0}} = \left[\frac{q}{C_{ox}(V_{GS} - Vth)} + \alpha\right] \cdot \Delta N_{nbti+hci}, \tag{4.9}$$

where $N_{nbti+hci}$ is the total trap density generated by NBTI and HCI.

4.3 I_{DD}-Based Aging Sensor Circuit Design

Subsequently, we describe the circuit architecture of the sensor that enables the proposed direct measurement of the NBTI and HCI induced aging effects. The block diagram of V_{th}-based and proposed I_{pp}-based aging measurement schemes are depicted in Figure 4.3. Different from existing V_{th}-based measuring scheme, our scheme measures the power supply current I_p directly from the CUO. The sensor consists of a Built-In Current Sensor (BICS), which mirrors the transient I_p current of the CUO, and sends it to a Current-mode Peak Detector (CPD). The CPD detects the peak value of the input current by using a current comparator, and holds the peak current for an adjustable time within a current memory, which allows the Current-To-Time converter (C2T) to translate the current value into a Pulse-Width Modulated (PWM) signal. With the PWM signal, the aging status of the CUO can be extracted by further processing with the model discussed in Section 4.2. This aging information can be further utilized to implement a DRM system, which can provide the best system performance for certain given application and reliability requirements.

The key component of the proposed sensor is the CPD, which is described in detail in the following section. In order to achieve a good accuracy, the C2T is carefully designed as well, its details being presented in Section 4.3.2. For the current sensing we make use of the BICS approach proposed in [77].

4.3.1 The Current Peak Detector (CPD)

Figure 4.4 depicts the proposed current peak detector, which is composed of: (i) a current memory cell with an adjustable memory holding time constant, to retain the peak current value, and (ii) a current comparator to determine if the present supply current value is bigger than the stored peak value, in which case the peak current value in the memory cell is updated. We denote by I_p, the input current of the CPD, by I_{pm} and I_{pp}, the input and output current of the current memory cell, and by V_{ctrl}, the output voltage of the current comparator. The current comparator [78], compares the values of I_p and I_{pm} currents. When the comparator input current is positive, i.e., $I_p > I_{pm}$,

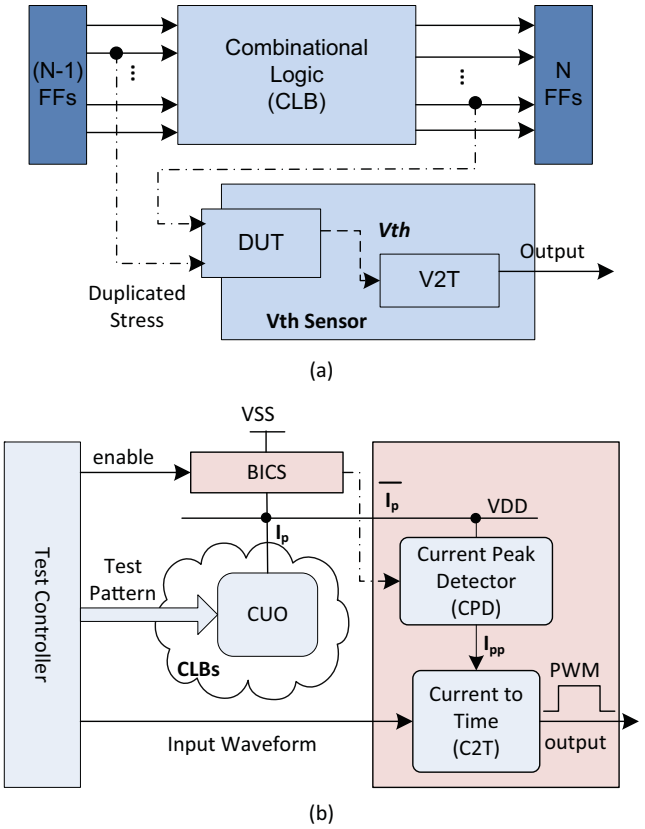


Figure 4.3: Two Different Measurement Schemes for Degradation Detection: (a) V_{th} Sensor Scheme; (b) A Direct Measurement Scheme with the Proposed I_{pp} Sensor (in the Shadow Box). The Aging Indicator I_{pp} of the Proposed Sensor Is Taken from the CUO Directly. V_{th} Sensor Takes the Aging Indicator V_{th} from the DUT of the Sensor.

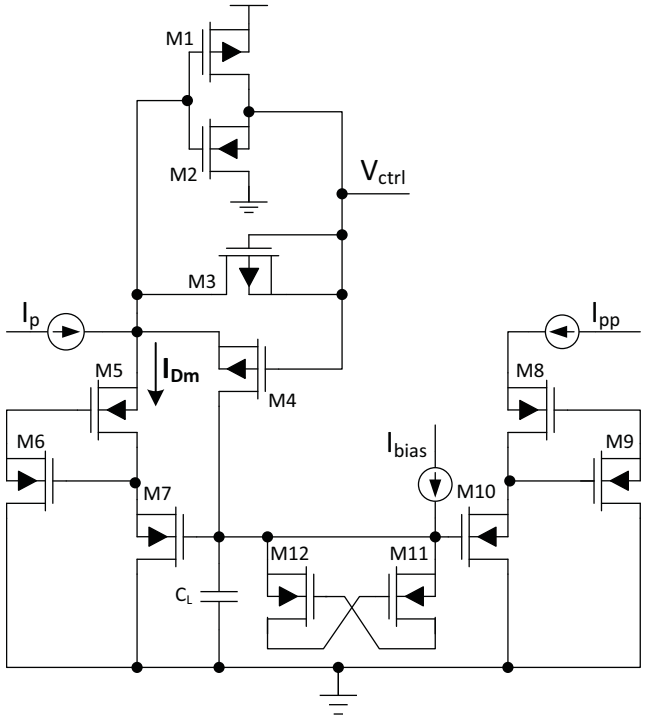


Figure 4.4: Circuit Schematic of the Current-Mode Peak Detector.

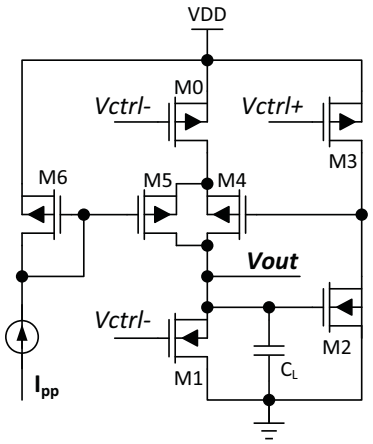


Figure 4.5: Circuit Schematic of the Current-to-Time Converter.

the comparator input voltage increases to the upper rail voltage V_{DD}, forcing the output voltage V_{ctrl} to be low. The current memory cell consists of two regulated cascode stages [79], which enable a better output swing and an increased output impedance. To describe the CPD operation, we distinguish between two functional stages: (i) the mirroring stage, characterized by $I_p > I_{pm}$ (i.e., the CPD input current is bigger than the stored peak value), and (ii) the peak holding stage, characterized by $I_p < I_{pm}$ (i.e., the CPD input current is smaller than the stored peak value, hence the peak value remains unchanged). During the mirroring stage, the comparator output voltage V_{ctrl} is low, enabling the current memory cell via transistor M_4, to mirror I_p (i.e., $I_{pp} = I_{pm} = I_p$). During the peak holding stage, V_{ctrl} is high, transistor M_4 is off, and the current memory cell holds the peak value of the I_p current (i.e., $I_{pp} = I_{pm}$). The storing capability is achieved using the gate-to-source capacitances of M_7 and M_{10}, in parallel with the discharge path consisting of cross coupled transistors M_{11} and M_{12} and current source I_{bias}. In this way, the discharging time constant can be controlled by the current I_{bias}. The maximum peak current is limited by the sourcing ability of the comparator with feedback diode, while the minimum detected peak current is constrained by the comparator gain and the output impedance of the circuit under aging assessment.

4.3.2 The Current-to-Time Converter (C2T)

The C2T converter, based on a thyristor delay element [80], is depicted in Figure 4.5. It receives as triggering input the comparator output voltage V_{ctrl} and generates a time interval proportional to the value of the control current I_{pp}. Its operation can be described as follows: When a rising edge of V_{ctrl} is detected, the load capacitance charges from 0 to V_{DD}, slower until the voltage V_{out} reaches the threshold voltage of transistor M_2, and faster for the remaining time to V_{DD}, due to the feedback connection. The falling edge of V_{ctrl} triggers the discharging of the voltage across capacitor

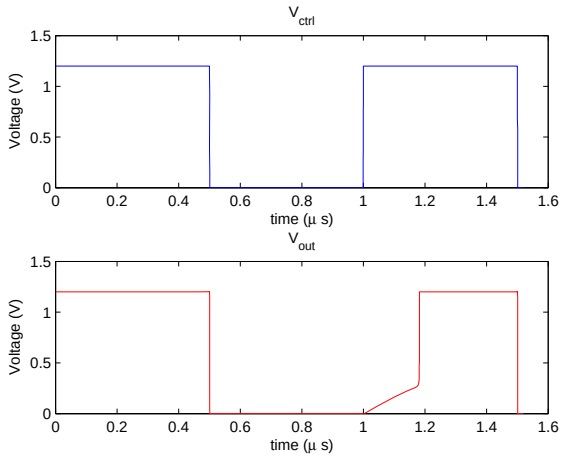


Figure 4.6: The Transient Waveform of the Voltage Across Capacitor C_L.

C_L through the drain-source capacitance of M_1 and the gate-source capacitance of M_2. The transient waveform of the voltage across the load capacitance C_L is presented in Figure 4.6. The C2T time delay is defined as the rising time of the converter voltage output and is given by the relation:

$$T = C_L \cdot V_{th_{M2}} / I_{pp},$$

where $V_{th_{M2}}$ denotes the threshold voltage of transistor M_2.

4.4 Performance Evaluation

The proposed aging sensor was implemented by using TSMC 65nm CMOS technology to analyze its performance. Figure 4.7 depicts the transient waveform of the I_p current and its afferent peaks, at a frequency of 1GHz.

In order to assess the accuracy of the peak detector and current-to-time converter circuits, we use a two stage operational amplifier as test vehicle. The reliability analysis of NBTI and HCI aging is carried out using Cadence RelXpert and Virtuoso Spectre simulators [55]. Figure 4.8 presents current-to-time converting results and the error evaluation of the peak detector circuit. The left axis represents the variation of the time delay T as a function of the control current I_{peak} for a load capacitance $C_1 = 1$pF. The right axis represents the measured peak value of I_p compared with the ideal peak value. For the purpose of illustration, we use a control current in the range 100uA $\sim$ 1mA, which results in a delay range of 120ns $\sim$ 40ns. Simulation results reveal that a fairly good linearity and accuracy are achieved.

In order to validate and evaluate the feasibility of our proposal, i.e., using the peak

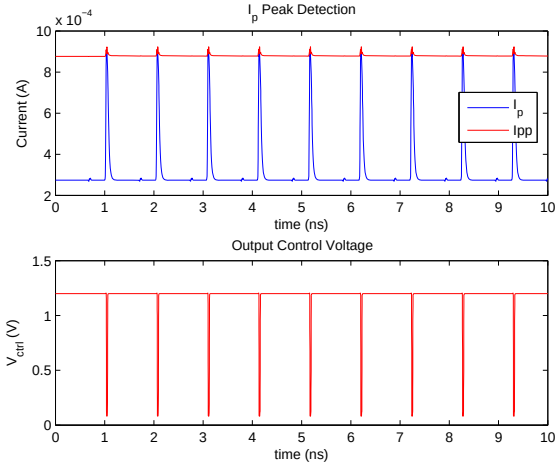


Figure 4.7: Peak Current Detection at 1GHZ.

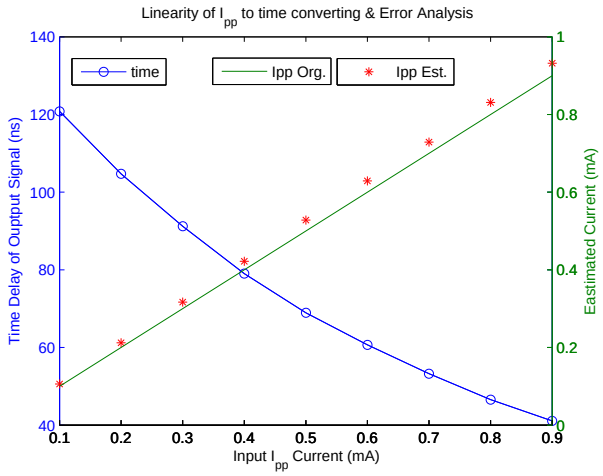


Figure 4.8: Linearity of Peak I_p to Time Converting (Left Axis) and Error Analysis of Peak Detection (Right Axis).

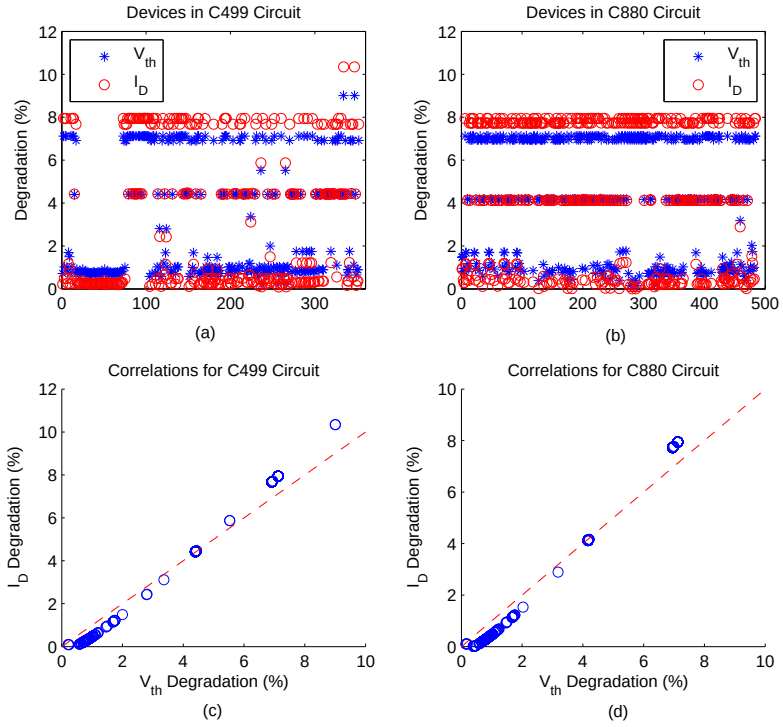


Figure 4.9: The Percentage Degradations of V_{th} and I_D for All Devices in the c499 and c880 Circuits - (a) and (b); and the Correlations between the Percentage Degradations of V_{th} and I_D - (c) and (d).

value of the I_p current as circuit aging monitor, we conducted accelerated testing simulation on the following ISCAS-85 benchmark circuits: c499, which is a 32-bit single error correcting circuit comprising 202 gates and c880, which is an 8-bit ALU, comprising 383 gates. The benchmark circuits are synthesized using the standard cells from TSMC 65nm technology library. The reliability analysis is carried by using Cadence RelXpert and Virtuoso Spectre simulators [55]. As concerns the simulation environment, we employed several input aggression profiles consisting of different input patterns for each benchmark circuit. As environment parameters, we used a temperature of $27^{\circ}C$, and a power supply $V_{DD} = 1.0$V. We exposed the benchmark circuits to NBTI/PBTI and HCI wearout stress and adopted an EOL target of 10 years. For each benchmark circuit, we determined its critical path. Then we measured the percentage degradation of the V_{th} and the drain current I_D for every transistor on the critical path.

The percentage degradation of V_{th} and I_D for all devices in the c499 and c880 circuits are graphically illustrated in Figure 4.9 (a) and (b). In can be observed that for both

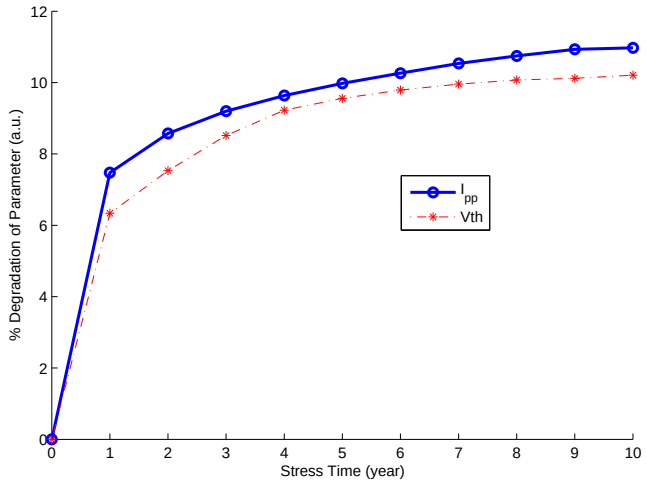


Figure 4.10: The Time Evolution of the V_{th} and I_D Degradation for 10-Year Simulation.

considered circuits, for those devices which are less degraded (i.e., the percentage of degradation is small), the I_D degradation is smaller than the V_{th} degradation. As the degradation percentage becomes larger, the I_D degradation value increases faster than the V_{th} value and eventually, towards the conventional EOL (i.e., 10% degradation of circuit critical parameters), it becomes larger than the V_{th} value. The improved sensitivity can be attributed to the dependence of I_D on multiple aging critical parameters, such as the threshold voltage V_{th} and the mobility μ. This means that the I_D degradation could be a better indicator than the V_{th} degradation is, if we are concerned with the reliability status of the IC more near its final operating stage. Figure 4.9 (c) and (d) depict the correlations between the percentage degradations of V_{th} and I_D. As it can be seen from the figures, the percentage degradation of I_D and V_{th} are strongly correlated, which means that both of them indicate the same aging trend for all the devices.

4.5 Conclusion

In this chapter, we proposed a direct aging measurement scheme and a novel aging sensor able to capture the amalgamated effect of NBTI and HCI degradation mechanisms on an IC "health" status. The sensor monitors the degradation of the peak power supply current in the Circuit Under Observation (CUO), enabling the measurement of the real aging status inside a circuit. We implemented the I_{pp}-based aging sensor in TSMC 65nm CMOS technology. The validity of using the I_D value to monitor the circuit aging, was analyzed and experimentally verified by means of simulation for a set of ISCAS-85 benchmark circuits. Compared to conventional V_{th} aging monitor,

I_D exhibits a similar aging rate for most of circuit lifetime. However, towards circuit EOL, the peak power supply current and the threshold voltage aging indications start to deviate from each other. Unfortunately, no other aging reference is available thus we cannot tell which method provides better estimates for that circuit lifetime part. We note however that high accuracy is essential during an IC early life, when it undergoes the greatest wear-out induced degradation, demanding thus accurate and fast reliability assessment and EOL prediction for an appropriate degradation compensation/mitigation measure.

Having addressed the aging sensors design, the next step is to assess the circuit reliability based on the sensors accrued aging status. In the next chapter we present such a model that can estimate a circuit end-of-life and its reliability statistics, while taking into account the sensors measurements uncertainty.

Note. The content of this chapter is based on the following papers:

N. Cucu Laurenciu, Y. Wang, S. D. Cotofana, **A Direct Measurement Scheme of Amalgamated Aging Effects with Novel On-Chip Sensor**, Proceedings of the 21^{st} IFIP/IEEE International Conference on Very Large Scale Integration (VLSI-SoC), pp. 246–251 (2013).

5

End-Of-Life and Reliability Statistics Estimation From Noisy Sensors Observations

For deca-nanometer advanced technology nodes, accurate reliability estimation is regarded as a key component of a dynamic reliability management framework as it can enable effective lifetime management via adopting appropriate mission profile specific policies. This chapter proposes an End-of-Life (EOL) prediction and reliability estimation approach, which takes into account the nonlinearities of the degradation process, as well as the sensors measurements and degradation process uncertainty, aiming to more realistically characterize the devices aging dynamics. Based on the degradation history, the estimation results are updated adaptively via the Bayesian method, once new degradation measurement data are provided. In order to validate and assess the estimation accuracy of the proposed framework, numerical simulations were performed on a power law degradation model. The obtained results for the considered nonlinear degradation process, reveal that, when compared with the commonly employed Wiener processes with linear mean, our approach exhibits improved estimation accuracy. Thus, it may be better suited to capture the nonlinearity and variability of in-field degradation dynamics and further to assess the actual circuit degradation status.

5.1 Introduction

The aggressive technology scaling for performance improvement has negatively impacted the devices lifetime reliability [81]. To address the aging-induced lifetime degradation of devices with minimal impact on the performance characteristics, Dynamic Reliability Management (DRM) frameworks have been developed. Since the effectiveness of the DRM policies depends on the reliability assessment accuracy, increasing attention has been payed to this topic. Most of past approaches [82–84] accept the modeling simplifying assumptions that the degradation process is mono-

ton, and/or can be linearized using time-scale transformations, which can result in a conservative lifetime estimation. Only recently, degradation models that integrate a nonlinear structure to trace better the degradation dynamics have been proposed. In [85], the mean parameter of the degradation process was updated using Kalman filtering, but its uncertainty was not considered and the variance was assumed linear in time. In [86], the degradation nonlinearity was captured without data transformations; however only the current degradation data was used, disregarding the degradation history. The degradation history problem was addressed in [87], but the variance was also assumed linear as in [58].

In view of the above, we propose a Bayesian reliability assessment framework, which takes into account the nonlinearities of the degradation process, aiming to characterize more realistically the wearout process dynamics and thus to improve the potential effectiveness of adopted reliability management policies. This is achieved by using a Wiener process to govern the dynamics of the degradation process, with nonlinear mean, which is expressed as a combination of basis functions, weighted by degradation history dependent parameters. Based on the entire degradation history, and not only on the instantaneous degradation state, the degradation model parameters are updated via Bayesian inference, once new degradation data are accumulated. Furthermore, we account in the proposed reliability assessment framework for the uncertainty in both the degradation process and the measurements. Numerical simulations were carried out in order to validate and evaluate the estimation accuracy of the proposed approach in comparison with commonly employed Wiener processes with linear mean. The obtained results quantitatively confirm that, when compared to the linear mean Wiener process, the proposed framework may be better suited for capturing nonlinear in-field degradation dynamics and hence for assessing the reliability and predicting the EOL in a more realistic manner.

The remaining of this chapter is organized as follows: Section 5.2 presents the degradation process formalism and the general modeling principles. The proposed framework is introduced in Section 5.3 and validated and evaluated in Section 5.4. Section 6.5 concludes the chapter with a summary of this work.

5.2 Degradation Process Formalism

Given that an Integrated Circuit (IC) is functional at the current time moment, based on its history of degradation (constituted by a set of noisy measurements collected from the in-field degradation sensors), one is interested in deriving its real-time reliability. According to the reliability status and the remaining operational life, appropriate lifetime management strategies can be adopted. Hence, the central problem of the IC reliability evaluation, is inferring the End-of-Life (EOL) statistics.

A degradation (wearout, aging) process is stochastic in nature. One candidate stochastic process that can govern the dynamics of an IC wearout process is the Wiener process, denoted subsequently by $W(t)$. The degradation process can be governed by an

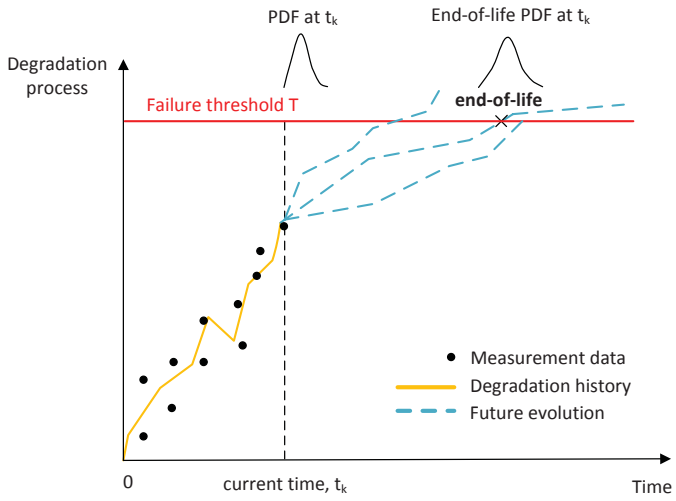


Figure 5.1: Illustration of the Reliability Modeling Principle.

equation of the form:

$$dX(t) = \mu(\alpha, t)\, dt + \sqrt{\sigma}\, dW(t), \tag{5.1}$$

where $X(t)$ describes the degradation state at time moment t. The Wiener degradation process $W(t)$ is specified by its mean (drift) μ, and variance $\sqrt{\sigma}$, which describe the degradation evolution in time. The nonlinearity of the degradation process is captured in the nonlinear time variation of the functional μ, with the parameters vector α. In order to accommodate for the heterogeneity of an IC degradation sources during its lifetime, the drift μ can be regarded as being composed of two terms: (i) $g(x, t)$, which is a fixed, deterministic component, common to all ICs (e.g., measurement bias), and (ii) $\alpha \cdot f(x)$, which is a variable, a-priori unknown nonlinear component, with $f(x)$, the set of basis functions (e.g., Gaussian, polynomial, fuzzy membership functions) and α, the unknown parameter vector. Consequently, (5.1) becomes:

$$dX(t) = g(x, t)\, dt + \alpha\, f(x, t)\, dt + \sqrt{\sigma}\, dW(t). \tag{5.2}$$

Therefore, the unknown parameters vector $\theta = (\alpha, \sigma)$ completely defines the degradation process, and has to be estimated from a set of noisy degradation measurements, $V(t)$. Having determined the IC degradation model, the future evolution of the degradation process can be predicted and the lifetime related properties of interest can thus be inferred. The general principle of the reliability estimation is graphically caught in Figure 5.1.

Given a set of noisy degradation measurements V (e.g., degradation of an IC performance characteristic such as max. operating frequency), which constitute the degradation history up to current time moment t_k, the degradation process parameters θ are

estimated. Based on the relation between a future degradation value and the up-to-date degradation history, given by the degradation process model, the potential future evolution paths of the degradation can be predicted. When a future degradation value exceeds a pre-specified threshold T (e.g., usually set to 10% degradation of the IC performance characteristic) for the first time, then the IC has reached its EOL. Hence, the EOL for a degradation path X can be defined as follows:

$$EOL = inf\left\{t : X(t) \geq T \mid X(s) < T,\ 0 < s < t\right\}. \tag{5.3}$$

The reliability at a time moment t for the ensemble of predicted degradation evolution paths, can then be obtained as the probability at time t of not reaching the EOL.

In view of the above, we shall present first the general methodology for deriving the device EOL in Section 5.3.1, followed by the corresponding algorithmic details in Sections 5.3.2 and 5.3.3.

5.3 Reliability Assessment Framework

For a given observation vector V, the parameters θ, which characterize the degradation process, are estimated taking into consideration the degradation history. The posterior distribution of the parameters θ is updated via a Bayesian framework [88], which enables to effectively integrate the historical, up-to-date degradation data together with the newly in-situ degradation observations. Once θ and the degradation path are estimated, the EOL is given by the time moment when the degradation path exceeds the predefined threshold. By simulating an ensemble of degradation paths for the same θ, the reliability at a specific time can be derived as the probability of not exceeding the predefined threshold.

5.3.1 Reliability evaluation procedure outline

The joint posterior distribution of θ and $x_{1:N}$, conditional on the observations V, can be sampled without having to compute the density, by using the Gibbs sampling algorithm [89], which alternates between the following two steps, for M times:

1. draw $x_{1:N} \mid \theta, V$, i.e., generate a sample of the degradation path $x_{1:N}$, for fixed θ and given observations V, and
2. draw $\theta \mid x_{1:N}, V$, i.e., generate a sample of the parameters set $\theta = (\alpha, \sigma)$, for fixed degradation path $x_{1:N}$ and given observations V.

The latter step, i.e., the update of the $\theta = (\alpha, \sigma)$ parameters, is particularly straightforward, since conjugate prior distributions are employed for α and σ. This makes it possible to derive analytically the conditional distribution of θ, given the observations V and the degradation path $x_{1:N}$, and sample directly from it. The former step

however, is more computationally demanding. In such a case, to overcome the difficulties of direct sampling, a Metropolis Hastings numerical approach [88] is applied. Specifically, the m-th iteration of the Gibbs sampler can be written as follows:

$$\begin{aligned}
&\text{sample } x_n^{(m)} \sim p\left(x_n \,|\, \theta,\ V,\ x_{n-1}^{(m)},\ x_{n+1}^{(m-1)}\right) \text{ for } n = 1 : N \\
&\text{sample } \sigma^{(m)} \sim p\left(\sigma \,|\, x_{1:N}^{(m)}, \alpha^{(m-1)}\right) \\
&\qquad\quad \alpha^{(m)} \sim p\left(\alpha \,|\, x_{1:N}^{(m)}, \sigma^{(m-1)}\right),
\end{aligned}$$

where $m = 1 : M$.
For the given observation vector V, having generated M samples of θ from the updated posterior distribution of $\theta \,|\, x_{1:N},\, V$, the EOL can now be inferred by simulating the M degradation paths using the discrete version of the continuous time dynamics governed by (5.2). For this purpose, we employ the Euler-Maruyama approximation [90], with a discretization time step of resolution τ, as follows:

$$x_{N+(k+1)\tau}^{(m)} = x_{N+k\tau}^{(m)} + \tau \cdot g(x_{N+k\tau}^{(m)}) + \tau \cdot \alpha^{(m)} f(x_{N+k\tau}^{(m)}) + \sqrt{\tau} \cdot Z_{N+k\tau}, \qquad (5.4)$$

where $Z_{N+k\tau} \sim \mathcal{N}(0, \sigma^{(m)})$. The number of discretization steps for each of the M paths, is determined by the EOL stopping criterion, i.e., when the degradation path sample $x_{N+(k+1)\tau}^{(m)}$ exceeds the EOL target (the threshold T, as defined in (5.3)). The EOL values for the M simulated paths, given the degradation history $x_{1:N}^{(m)}$ and the corresponding parameters set $\theta^{(m)}$ for each path, are computed as:

$$\mathbf{EOL}_{1:M} = inf\left\{EOL_m \,:\, x_{EOL_m}^{(m)} \geq T,\ \ m = 1 \ldots M\right\}. \qquad (5.5)$$

The reliability function at time instant t, can now be derived as:

$$R(t) = \{P(\mathbf{EOL}_{1:M} > t) \,:\, t > N\}. \qquad (5.6)$$

With the above considerations in place, we are now in position to present the Gibbs sampler details for obtaining the parameters which characterize the degradation process.

5.3.2 Gibbs sampler step 1 - draw $x_{1:N} \,|\, \theta, V$

Given the parameters set $\theta = (\alpha, \sigma)$, the distribution of a degradation path $x_{1:N}$ can be obtained from:

$$p(x_{1:N} \,|\, \theta, V) \propto p(V \,|\, x_{1:N})\, p(x_{1:N} \,|\, \theta), \qquad (5.7)$$

where $p(x_{1:N} \,|\, \theta)$ is the probability of deriving the degradation path $x_{1:N}$ for the parameters set θ, and $p(V \,|\, x_{1:N})$ is the likelihood, the probability of observing the measured degradation path from $x_{1:N}$. The Markov property [88] of (5.4) implies that the conditional distribution of $x_n^{(m)}$, given all the other values is the same as the distribution

given the adjacent endpoints $x_{n-1}^{(m)}$ and $x_{n+1}^{(m-1)}$. It follows that the posterior distributions of the parameters are:

$$p(x_n^{(m)} \mid x_1^{(m)}, \dots, x_{n-1}^{(m)}, x_{n+1}^{(m-1)}, \dots, x_N^{(m-1)}, \theta, V) \propto p(x_n^{(m)} \mid x_{n-1}^{(m)}, x_{n+1}^{(m-1)}, \theta, V),$$

which is further proportional to:

$$\propto p(V_n \mid x_n^{(m)})\, p(x_{n-1}^{(m)} \mid x_n^{(m)}, \theta)\, p(x_{n+1}^{(m-1)} \mid x_n^{(m)}, \theta). \tag{5.8}$$

As concerns $p(V_n \mid x_n^{(m)})$, it results from the distribution of the noisy observations, which is given a-priori. Since $x_n \mid x_{n-1}$ in (5.4) follow a normal distribution, the probabilities $p(x_{n-1}^{(m)} \mid x_n^{(m)}, \theta)$ and $p(x_{n+1}^{(m-1)} \mid x_n^{(m)}, \theta)$ can be readily derived.

For the present purposes, in order to sample the target conditional distribution of $x_{1:n} \mid \theta, V$, we employ a numerical procedure, i.e., the Metropolis Hastings algorithm [88], outlined subsequently. Based on (5.4) with time step resolution h (which can be equal to the time sampling resolution of observed data), for existing x_n, a step x_n^* can be proposed by drawing from the distribution:

$$pdf(x_n^* \mid x_n) = \frac{1}{\sqrt{h\sigma 2\pi}} \cdot e^{-\frac{(x_n^* - x_n - h\,g(x_n) - h\,\alpha f(x_n))}{2h\sigma}}. \tag{5.9}$$

The acceptance probability of x_n^* as candidate to replace the current draw x_n, is given by:

$$\rho\left(x_n^{(m)}, x_n^{*(m)}\right) = \min\left\{\frac{q(x_n^*, x_n)}{q(x_n, x_n^*)}, 1\right\}, \tag{5.10}$$

$$\text{where } q(x_n^*, x_n) = p(V_n \mid x_n^*)\; p(x_{n+1} \mid x_n^*)\; p(x_n^* \mid x_{n-1})$$
$$q(x_n, x_n^*) = p(V_n \mid x_n)\; p(x_{n+1} \mid x_n)\; p(x_n \mid x_{n-1}).$$

The conditional probabilities in (5.10) can be computed using (5.9).

5.3.3 Gibbs sampler step 2 - draw $\theta \mid x_{1:N}$

We are interested in drawing a sample of the parameters θ from the posterior probability distribution, which is given by:

$$p(\theta \mid x_{1:N}) \propto p(\theta)\; p(x_{1:N} \mid \theta). \tag{5.11}$$

To this end, we derive the likelihood of θ for a given degradation path $x_{1:N}$ as:

$$p(x_{1:N} \mid \theta) \propto \frac{1}{\sqrt{\sigma^N}} \cdot e^{-\frac{1}{2h\sigma} \cdot \sum_{n=1}^{N} [x_{n+1} - x_n - h \cdot g(x_n) - h \cdot \alpha f(x_n)]^2}. \tag{5.12}$$

For a successful Bayesian inference, we assume the prior distributions for the parameter set $\theta = (\alpha, \sigma)$ belong to the conjugate family [88] of the sampling distribution $p(x_{1:N} \mid \theta, V)$. As such, we consider an inverse Gamma distribution for the degradation process variance, i.e., $\sigma \sim \Gamma^{-1}(q_2, q_3)$, and a normal distribution for α, i.e., $\alpha \sim \mathcal{N}(0, q_1)$. Since

$$\begin{aligned} p(\alpha \mid x_{1:N}, \sigma) &\propto p(\alpha)\, p(x_{1:N} \mid \alpha, \sigma) \text{ for } \sigma \text{ known} \\ p(\sigma \mid x_{1:N}, \alpha) &\propto p(\sigma)\, p(x_{1:N} \mid \alpha, \sigma) \text{ for } \alpha \text{ known,} \end{aligned}$$

it follows that:

$$\begin{aligned} \alpha \mid x_{1:N}, \sigma &\sim \mathcal{N}(\alpha_m, \alpha_v) \\ \alpha_m &= -\alpha_v \cdot \sigma^{-1} \sum \left[x_{n+1} - x_n - h \cdot g(x_n)\right] \cdot f(x_n) \\ \alpha_v &= \left[q_1^{-1} + \sigma^{-1} \cdot h \cdot \sum f^2(x_n)\right]^{-1} \\ \sigma \mid x_{1:N}, \alpha &\sim \Gamma^{-1}(\sigma_m, \sigma_v) \\ \sigma_m &= q_3 + N \\ \sigma_v &= q_2 + h^{-1} \sum \left[x_{n+1} - x_n - h \cdot g(x_n) - h \cdot \alpha f(x_n)\right]^2 . \end{aligned}$$

5.4 Performance Evaluation

In order to validate the proposed approach, we consider the nonlinear process modeled by (5.2), with mean $\mu(\alpha, t)$ given by t^β ($g(x, t) = 0$). We conduct the numerical experiments employing the following parameters values: the number of degradation paths equal to 100, the discretization step $h = 0.1$, the Wiener process variance $\sqrt{\sigma} = 0.23$, and its mean power-law coefficient $\beta = 2$. As concerns the basis functions α modeling the process mean, without loss of generality, for simulation purposes, we employed the Gaussian kernel [91]. The estimation performance of the proposed model was studied using noisy observations sampled from $\mathcal{N}(x(t), 0.01)$. For estimation accuracy evaluation purpose, we compare with the commonly employed Wiener processes with linear mean given by $\alpha \cdot t$ [82–84].

In Figure 5.2, the real degradation path generated via (5.4), is illustrated against the two degradation paths, estimated with the proposed nonlinear degradation model and with the reference linear model, respectively. In direct relation to ICs aging, the degradation path could represent the threshold voltage degradation of a transistor, the maximum operating frequency degradation of circuit, etc. It can be observed that, the proposed nonlinear degradation model exhibits a fairly well estimation ability, the real and estimated degradation paths almost overlapping. Furthermore, the fitting characteristics are improved when compared to the reference model with linear drift, for the considered degradation process. We note that while the estimation tends to be less accurate during the early period, characterized by a limited degradation history, the fitting characteristics improve in time, as degradation data are accumulated. The

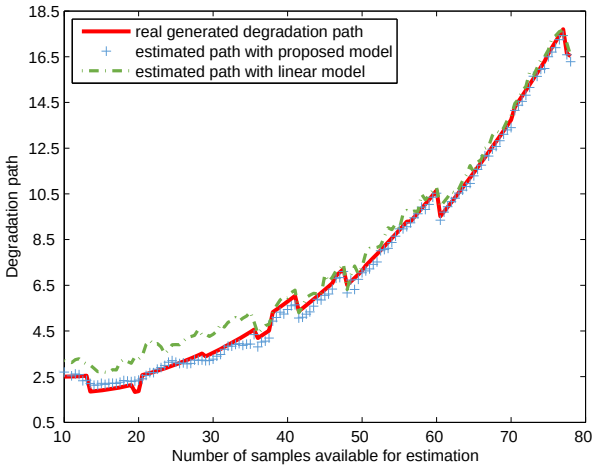


Figure 5.2: The Real vs. Estimated Degradation Paths.

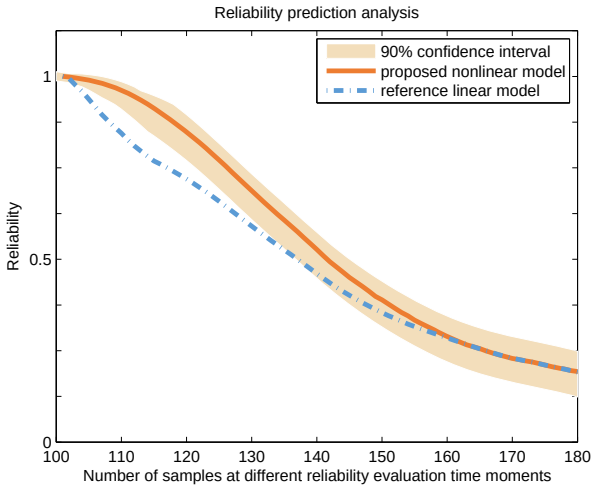


Figure 5.3: The Reliability Assessment/Prediction Error Analysis.

proposed model however yields a more accurate estimation than the model with linear mean, also during the early period with few degradation measurement data. We evaluated the degradation path estimation accuracy using the Akaike Information Criterium [92], whose statistics are given by:

$$AIC = 2 \cdot k - 2 \cdot \max\left[\log\left(likelihood\right)\right], \tag{5.13}$$

and which provides a measure of the trade-off between the model's complexity (reflected in the number of estimated parameters, k) and the goodness-of-fit (reflected in the log-likelihood of the fit). The better estimation accuracy using the proposed model is quantitatively reflected in a smaller AIC value (121), when compared to 135 achieved by the linear model.

Based on obtained estimates, the in-field reliability is evaluated at different time moments, as depicted in Figure 5.3. In order to address the reliability assessment uncertainty, we derived the confidence interval using the bootstrap method [93]. Simulation results reveal that not being able to capture accurately the degradation process nonlinearities, can result in an underestimation of the reliability, especially during the initial degradation period. This is turn may yield a less efficient lifetime management of the device whose reliability is being assessed. In the ICs case, for instance, being able to realistically asses the reliability status during the early in-field period, is of particular interest. This is because of the front-loaded nature (i.e., the highest extent of degradation is manifested during early operation, after which the degradation tends to saturate) of the front-end-of-life aging mechanisms, which imply that the lifetime management strategies are most effective during early life.

5.4.1 Case Study

For expository purposes of proposed approach assessment of validity and potential applicability in reliability management frameworks, we provide subsequently a practical case study using the aging data of a PMOS transistor. To this end, we conducted accelerated aging simulation of a PMOS transistor implemented in PTM 45nm technology [73]. As aging quantifier we employ the transistor threshold voltage, V_{th} [27]. The reliability analysis is carried in Cadence RelXpert and Virtuoso Spectre simulators [55], using the substrate and gate current, lifetime and AgeMos model parameters extracted in BSIMPro+ [72] for the PTM 45nm technology. As concerns the environmental parameters, we used a temperature of 27°C, and a power supply VDD $= 1.0$V. We exposed the PMOS transistor to Negative Bias Temperature Instability (NBTI) and Hot Carrier Injection (HCI) wearout stress and adopted an EOL target (failure threshold T) of 9 years.

The percentage degradation of the transistor V_{th} is graphically illustrated in Figure 5.4. The V_{th} time evolution, as obtained from Cadence simulation, serves as the real degradation data. Based on the V_{th} data, the noisy observations are then obtained in a similar manner with the synthetic example previously studied, specifically by sampling from the distribution $\mathcal{N}(V_{th}(t), 0.01)$. We derived the transistor EOL values,

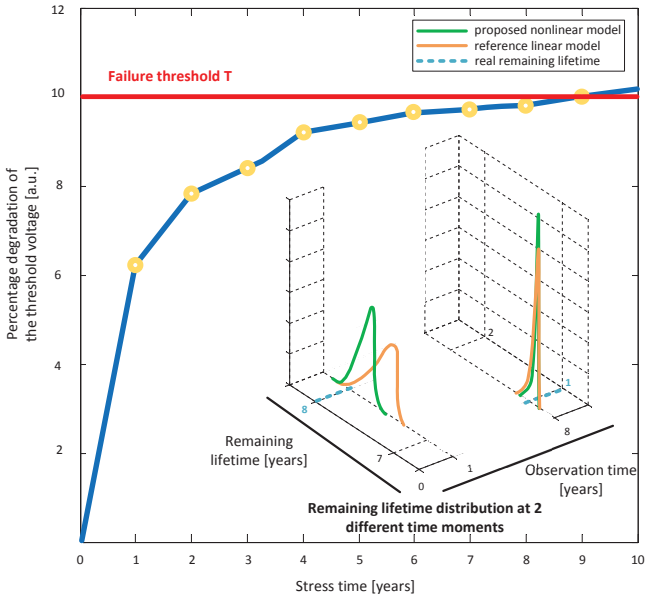


Figure 5.4: Time Evolution of the V_{th} Degradation After 10-year Simulation and Remaining Lifetime Distribution at two Different observation time moments: 1 year and 8 years.

using the proposed approach and the linear model approach, at two different observation time moments: 1 year and 8 years, respectively. Based on the EOL values, the transistor remaining lifetime values were then obtained, each as the difference between the EOL time moment and the current observation time moment. The corresponding Probability Density Functions (PDFs) of the remaining useful lifetime values estimated with both proposed and linear approach, and the real remaining lifetime values obtained from Cadence, are depicted in Figure 5.4 for comparison.

As it can be observed in Figure 5.4, at the beginning of the transistor operating life, the uncertainty in the estimated remaining lifetime PDFs, under both proposed and the linear approach, is higher. However, our model outperforms the linear counterpart, with a more precise estimation spread and a PDF mean value closer to the real transistor remaining lifetime value. The early EOL and implicitly the remaining lifetime estimation accuracy differences between the two approaches, can be attributed to the ability to capture the nonlinearities exhibited by the V_{th} degradation observations. As the circuit ages and more degradation observations become available, the EOL prediction uncertainty cones get narrower, and the differences between the two distributions become smaller.

When limited degradation observations are available, the accuracy of early EOL predictions is more sensitive to the selection of the prior distribution of $\theta = (\alpha, \sigma)$, which characterizes the degradation process, i.e., an inappropriate selection of these initial

parameters, causes the predictions to be less accurate with smaller confidence intervals. Such is the case in the considered simulation setup which yields less accurate EOL predictions both for our approach and for the reference linear one, during the transistor early life, as illustrated by the two PDFs in Figure 5.4 at 1 year observation time. However, the proposed approach takes into account the nonlinearities of the degradation process and is less sensitive to the selection of the prior distribution, exhibiting better adapting ability as far as the θ updating is concerned and, as a consequence, better prediction accuracy when compared to the linear model. Improved accuracy of EOL predictions during the early life stages, can be achieved if the prior distribution of $\theta = (\alpha, \sigma)$ parameters is restricted to meaningful values. However, for the current technology nodes with the afferent highly dynamic variability threats, precise knowledge based on experience with the same failure mechanisms in similar components may be harder to obtain.

As the amount of available degradation observations increases, the predictive ability improves for both approaches, as the posterior PDF becomes dominated by the likelihood given by (5.12), situation exemplified in Figure 5.4 by the two PDFs at 8 years observation time.

The previously studied practical case, illustrates the significance of incorporating nonlinearity in the degradation process model when the underlying process is nonlinear, especially when EOL predictions are desired during the beginning of the device life, characterized by limited degradation history.

5.5 Conclusion

In this chapter we proposed a Bayesian EOL prediction and reliability estimation framework that takes into account the degradation process nonlinearity and uncertainty, from noisy observations. Based on the degradation history and the current measurement data, the degradation process parameters are updated via the Bayesian method. As such, future degradation evolution can be derived and the afferent reliability statistics estimated. Simulation results revealed that the degradation process dynamics and the reliability evolution can is more accurately captured by the proposed approach, when compared with a linear mean process based counterpart. Moreover the estimation method is general and can be utilized for the EOL prediction of any circuit (block) whose aging status is being monitored by sensors.

Thus far we have addressed reliability assessment at the transistor/circuit level, which results can be utilized to guide: (i) the 4-dimensional (area, delay, power, reliability) design exploration process and (ii) the selection of the aging mitigation/compensation mechanisms to be included in the reliability wrapper, and at run time to guide the lifetime reliability management policy. A complementary line of reasoning is to embrace the inherently unreliable operation of transistors/circuits and investigate circuits/architectures that are able to perform computation in a reliable manner. The next chapter we pursue this view angle change and propose a coding based approach

towards reliable data transport via unreliable interconnects.

Note. The content of this chapter is based on the the following paper:

N. Cucu Laurenciu, S. D. Cotofana, **A Nonlinear Degradation Path Dependent End-of-Life Estimation Framework from Noisy Observations**, Microelectronics Reliability 53(9–11), pp. 1213–1217 (2013).

6

Reliable Data Transport

In this chapter we introduce and evaluate Haar based codec assisted medium and long range data transport structures, e.g., bus segments, Network on Chip interconnects, able to deal with technology scaling related phenomena (e.g., increased susceptibility to proximity coupling noise and transmission delay variability), targeting energy savings at the expense of a reasonably small overhead, i.e., 1 extra wire, a 2-gate encoder, and a 2-gate decoder, for each and every pair of uncoded wires. For practical evaluation we employed a 45nm commercial CMOS technology and different random, uncorrelated workload profiles.

For 5mm and 10mm long 8-bit buses (without repeaters), we obtain energy savings of 55% and 34%, and a transmission frequency increase of 35% and 41%, respectively, at the expense of less than 1% area overhead with respect to the reference system (i.e., 8-wire synchronous uncoded bus), which prove energy and delay effectiveness. We further augment our proposal with a Single Error Correction and Double Error Detection (SECDED) scheme particularly adapted to its structure, in order to cope with very deep sub-micron noise (e.g., supply voltage variations, electromagnetic interference) induced transmission errors. When compared to the reference system (not SECDED protected), for 10mm long buses, the Haar tailored SECDED approach consumes 27% less energy at the expense of 2% area overhead, while providing robuster data transmission.

6.1 Introduction

Driven by the global economy growth impetus manifested in the semiconductor industry as an escalating need for increased processing speed, quality, and yield, the CMOS scaling evolutionary trend has fueled the exponential rise of computing performance for the past 40 years, leading to higher IC integration density, and reduced gate delay, parasitic capacitance, energy and active power consumption per transition. Conversely, as the atomic/nanoscale size silicon limitations are approached, undesirable variability and stochastic error-prone device operation become prominent [94].

As wires are getting narrower and taller, and the wire pitch smaller, the interconnect parasitics worsen, the inter-wire capacitance increases (and thus the susceptibility to neighboring wires interference) negatively impacting the transmission latency and power consumption (the interconnects dominating figures [38]), as well as the transmitted signals integrity. In view of these considerations, in concert with logic optimization, a multi-criteria (e.g., energy, delay, physical area, reliability) design-time interconnect-centric avenue becomes a critical desideratum for high-performance and/or low power SoCs.

En route to address interconnect specific challenges, several solutions have been proposed, e.g., wire shielding and spacing, low-swing signaling, charge recycling, buffer/repeater insertion, and coding techniques [95]. The coding-based strand of research constitutes a technology and implementation independent compelling alternative, appealing from the power-delay-reliability multi-objective optimization standpoint. Several coding techniques have been previously investigated [95–99], most of them being focused on a single bus desideratum criterion (i.e., low power or low latency or reliability). Low power and low delay codes have as salient target switching activity reduction by taking advantage in the coding process of the temporal/spatial signature of the to be transmitted data. Most of these techniques are either tailored for data buses [100–102] or for address buses [103–108], and as a result whilst effective for specific buses data profile correlations and peculiarities, they are less suitable for a general transmission context with random, uncorrelated data. To combat and decrease interconnects susceptibility to the Very Deep Sub-Micron (VDSM) noise (e.g., supply voltage variations, electromagnetic interference) induced errors that can arise during the data transport, error control bus coding was employed [95]. However, as the error detection and correction codec is generally very computationally involved, its combination with low power or latency bus coding techniques has not been envisaged.

In light of the above perspective, in this paper we introduce and evaluate in a commercial 45nm technology codec assisted energy effective reliable data transport structures, e.g., bus segments, Network on Chip (NoC) interconnects, able to deal with technology scaling related phenomena, e.g., crosstalk and transmission delay variability, at the expense of a reasonably small area overhead. To this effect, we propose a low complexity $2 : 3$ single stage Haar Transform based codec, which enables energy savings while also alleviating data transport time related aspects (i.e., we diminish the data transmission latency and obtain a lower variability data arrival profile, which is a key issue for interconnect robustness (reliability) in the context of VDSM fabrication technologies high process parameter variability). Moreover, we augment the Haar-based codec with a Single Error Correction Double Error Detection (SECDED) capability adapted to its peculiarities, such that we can combat errors arising during the data transport process. To assess the practical implications of our proposal, an 8-bit wide interconnect segment is equipped with the proposed codec infrastructure and evaluated for varying interconnect length and width. Our simulations indicate that when compared to the reference uncoded interconnect our proposal enables 30%, 55%, and 34% energy savings for an interconnect length of 1mm, 5mm, and 10mm, respectively. Moreover, given that the considered data encoding schemes diminish the

crosstalk occurrence, codec augmented interconnects longer than 1mm can be operated at a higher frequency than the uncoded ones. In particular, a clock frequency increase of about 35% and 41% is enabled for an interconnect length of 5mm and 10mm, respectively. The energy and data transmission delay reductions are obtained at an area increase of less than 1% with respect to the reference uncoded design. The direct Error Correcting Code (ECC) augmentation of a 10mm long 8-wire reference system results in a $1.33\times$ energy increase, while for the same bus length the ECC enhanced Haar system requires 2% area overhead, consumes with 27% less energy and operates at a slightly higher frequency than the reference uncoded counterpart.

The remaining of the chapter is organized as follows: The Haar coding scheme algorithmic aspects are discussed in Section 6.2, together with a succinct presentation of the afferent encoder and decoder architecture. Section 6.3 deals with the evaluation of the proposed codec practical implications. Section 6.4 briefly reviews recent related work on bus coding and compares our proposal with existing state-of-the-art. Finally, Section 6.5 concludes the chapter.

6.2 Haar Codec Modus Operandi

For the nanometer technology, as the inter wire capacitances dominate the total bus capacitance, crosstalk between adjacent wires becomes a prominent concern. The increased capacitive coupling effects include glitches and/or increase of the transmission delay along the bus (when adjacent wires are switching in opposite directions, the transition on one wire might be slowed down), and in turn to the overall power consumption increase. One way to diminish these effects is to encode the transmited data, such that coupling transitions ("1" $\rightarrow$ "0" and viceversa) between adjacent wires are as scarce as possible. To this effect, subsequently, we: (i) introduce a codec suitable for medium and longer range interconnects, whose implementation is presented Section 6.2.1, that simultaneously targets energy, area, and delay merits, and (ii) augment the codec with error detection and correction circuitry particularly tailored to its structure, as described in Section 6.2.2. At the crux of our codec, lies a $2:3$ stage 1 Haar Transform [109] with a lightweight implementation, which takes advantage of bits compression benefits to reduce both the wires own transition count (less switching in time along each wire, as illustrated in Figure 6.1-a)), and coupling transition count (less occurrences of adjacent wires switching concomitantly in opposite directions, as illustrated in Figure 6.1-b)). For brevity, we assume that a byte-wise synchronous data transmission is desired but the discussion is general and can be easily extended for other interconnect widths.

6.2.1 Haar Codec

The Haar encoder receives as input per each clock cycle, a data byte subsequently denoted by $\{x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$, and generates as output a 12-bit wide encoded

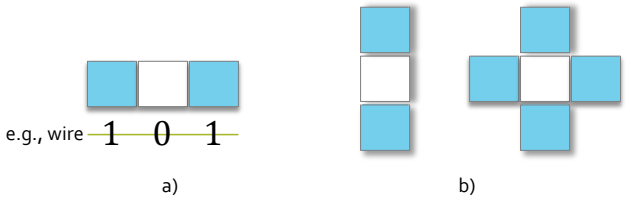


Figure 6.1: Switching patterns to be avoided for: a) own transitions along the wire, and b) single/double coupling transitions between adjacent wires.

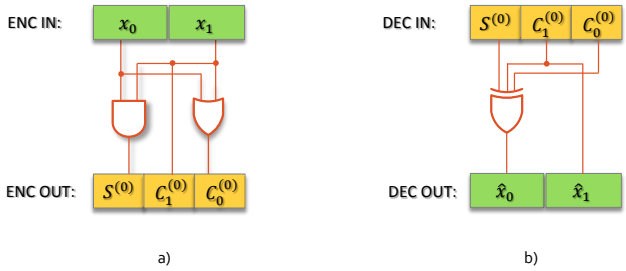


Figure 6.2: 2- bit Haar codec: a) encoder, and b) decoder.

sequence to be sent over the bus wires.
To this end, the input data sequence is divided into 4 2-bit groups as follows: $\{x_0, x_1\}, \{x_2, x_3\}, \{x_4, x_5\}, \{x_6, x_7\}$. For each such pair of input bits, the encoder performs 1-bit pair-wise addition and subtraction. Normally, the sum of two bits in $2's$ complement notation requires two bits for representation (as the exact sum value is required). However, because we compute both the sum and the difference of the same bits only the sum MSB (which corresponds to the carry out signal of a 1-bit full adder) needs to be evaluated. This is because the sum LSB is the parity bit, which is identical to the difference LSB (since the sum of two bits has the same parity as their difference). As concerns the difference of two input bits, it is likewise performed in $2's$ complement and requires a 2-bit representation. Specifically, for the input bits $\{x_0, x_1\}$, for instance, the encoder computes the following three bits:

$$\begin{aligned} S^{(0)} &= x_0 \wedge x_1 \\ C_1^{(0)} &= x_1 \\ C_0^{(0)} &= x_0 \vee x_1, \end{aligned}$$

where $\wedge$ denotes a logical AND operation, and $\vee$ denotes a logical OR operation, as illustrated in Figure 6.2-a).

The Haar decoder receives as input 12 bits of encoded data $\left\{S^{(0)}, C_1^{(0)}, C_0^{(0)}, S^{(1)}, C_1^{(1)}, C_0^{(1)}, \ldots, S^{(3)}, C_1^{(3)}, C_0^{(3)}\right\}$ and outputs the data byte

$\{\hat{x}_0, \hat{x}_1, \hat{x}_2, \hat{x}_3, \hat{x}_4, \hat{x}_5, \hat{x}_6, \hat{x}_7\}$. Conceptually speaking, if we know the sum and difference of two numbers, it is straightforward to compute the two numbers in cause. Exemplifying for the bit pair $\{x_0, x_1\}$, the following equations govern:

$$\hat{x}_0 = S^{(0)} \oplus C_1^{(0)} \oplus C_0^{(0)}$$
$$\hat{x}_1 = C_1^{(0)},$$

where $\oplus$ denotes a logical XOR operation, as illustrated in Figure 6.2-b) and for an error-free transmission $\hat{x}_0 = x_0$ and $\hat{x}_1 = x_1$.

The following architectural related observations are in order:
− *Codec Implementation Complexity, Delay, and Energy.*
As concerns the hardware implementation, the Haar encoder and decoder exhibit very low complexity, consisting of one logic level for the encoder (one OR/AND gate) and two logic levels for the decoder (two XOR gates for the decoder, respectively). As a result, the Haar encoder/decoder delay is very small, i.e., the delay of a single logic gate for the encoder and two logic gates the decoder. The codec simplicity has also positive implications on the energy consumed by the encoder/decoder.
− *Codec Scalability to Wider Interconnects.*
The low hardware complexity enables its utilisation for wider interconnects, as it scales linearly with respect to the number of wires (e.g., for 4 wires, the encoder requires 2 parallel OR gates and 2 parallel AND gates; for 8 wires, it requires 4 OR gates and 4 AND gates).

6.2.2 SECDED

For ECC data protection, two avenues can be followed, i.e., either protect the original bits to be transmitted over the bus, or the actual bits that are sent on the bus. Subsequently, we shall present the former approach algorithmic details, as it enables us to take advantage of the particular Haar codec structure for energy reduction.

Let m_0 denote the encoded sequence of transmitted bits $m_0 = (S^{(0)}, C_1^{(0)}, C_0^{(0)})$, and ϵ the transmission error pattern. Table 6.1 summarizes all possible 1-bit error scenarios affecting m_0. The first two columns in the table represent the original data bits x_0 and x_1, the third and the fourth column denote the Haar encoded message at the interconnect transmitting end, and the message at the receiving end (Haar encoded message + noise), respectively; while the last two columns correspond to the estimated data bits after Haar decoding. As for single bit errors, all possible values that ϵ can take are: (0 1 0), (1 0 0), and (0 0 1), 3 situations should be analyzed at the interconnect receiving end for each possible x_0 and x_1 bit combination. One may note in Table 6.1, that in all one-error scenarios, the decoded value $\hat{x}_0$ is always erroneous. This is expected, as all three encoded bits $(S^{(0)}, C_1^{(0)}, C_0^{(0)})$ are involved in the computation of the decoded bit $\hat{x}_0$. It follows that any single error affecting the encoded bit sequence $(S^{(0)}, C_1^{(0)}, C_0^{(0)})$, will always result in an erroneous $\hat{x}_0$ value. Thus it is mandatory to protect x_0 in order to be able to correct an erroneously decoded value $\hat{x}_0$. On the other

Table 6.1: One-Bit Error Scenarios for Haar System.

x_0	x_1	(m_0)	$(m_0 \oplus \epsilon)$	$\hat{x}_0$	$\hat{x}_1$
			(1 0 0)		1 (F)
0	0	(0 0 0)	(0 1 0)	1 (F)	1 (F)
			(0 0 1)		0
			(1 1 1)		0 (F)
0	1	(0 1 1)	(0 0 1)	1 (F)	0 (F)
			(0 1 0)		1
			(1 0 1)		1 (F)
1	0	(0 0 1)	(0 1 1)	0 (F)	1 (F)
			(0 0 0)		0
			(0 0 0)		0 (F)
1	1	(1 0 0)	(1 1 0)	0 (F)	0 (F)
			(1 0 1)		1

hand, when $\hat{x}_0$ is erroneous, we also need to be able to discriminate the correct value of $\hat{x}_1$, in which case $C_0^{(0)}$ has to be protected. Generalizing from the two input bits $\{x_0, x_1\}$ to the entire input byte $\{x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$, we propose to append the following 5 error control coding bits to the 12-bit Haar encoded sequence:

$$
\begin{aligned}
E_1 &= x_0 \oplus x_2 \oplus x_6 \\
E_2 &= x_0 \oplus x_2 \oplus x_4 \\
E_3 &= x_2 \oplus x_4 \oplus x_6 \\
E_4 &= x_0 \oplus x_4 \oplus x_6 \\
E_5 &= C_1^{(0)} \oplus C_1^{(1)} \oplus C_1^{(2)} \oplus C_1^{(3)}
\end{aligned}
$$

The $\{E_1, E_2, E_3, E_4\}$ bits correspond to a $(7, 4)$ Hamming code and are used for the correction of the input bits $\{x_0, x_2, x_4, x_6\}$, while bit E_5 is simply a parity bit used for the correction of the input bits $\{x_1, x_3, x_5, x_7\}$.

Single Error Detection and Correction
Any single bit-flip error affecting any bit of the 17-bit sequence

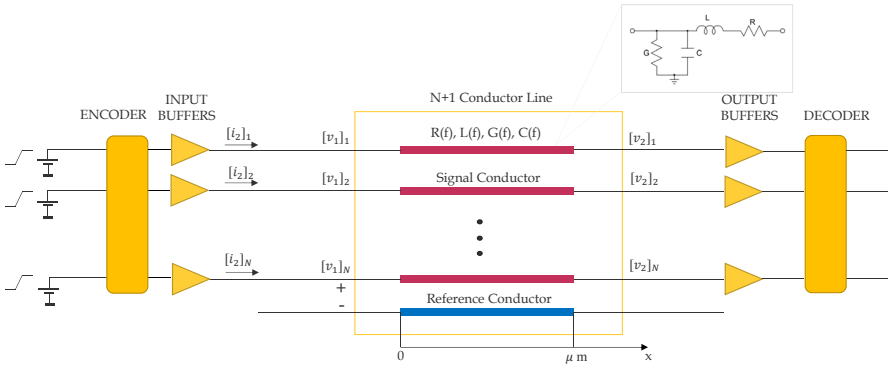


Figure 6.3: SPICE simulation setup for the interconnect coding-based systems.

Table 6.2: Haar System Single Error Correction.

Case 1	Case 2
$\hat{x}_0 = S^{(0)} \oplus C_1^{(0)} \oplus C_0^{(0)}$	$\hat{x}_0 = E_1 \oplus E_2 \oplus E_4$
$\hat{x}_2 = S^{(1)} \oplus C_1^{(1)} \oplus C_0^{(1)}$	$\hat{x}_2 = E_1 \oplus E_2 \oplus E_3$
$\hat{x}_4 = S^{(2)} \oplus C_1^{(2)} \oplus C_0^{(2)}$	$\hat{x}_4 = E_2 \oplus E_3 \oplus E_4$
$\hat{x}_6 = S^{(3)} \oplus C_1^{(3)} \oplus C_0^{(3)}$	$\hat{x}_6 = E_1 \oplus E_3 \oplus E_4$

$\left\{S^{(0)}, C_1^{(0)}, C_0^{(0)}, \ldots, S^{(3)}, C_1^{(3)}, C_0^{(3)}, E_1, \ldots, E_5\right\}$, can be corrected as follows: We compute in parallel each of the bits $\{\hat{x}_0, \hat{x}_2, \hat{x}_4, \hat{x}_6\}$ in two manners, as summarized in Table 6.2. The bits $\{w_1, w_2, w_3, w_4\}$ computed as $w_1 = \hat{x}_{0\ (\text{case 1})} \oplus \hat{x}_{0\ (\text{case 2})}$, $w_2 = \hat{x}_{2\ (\text{case 1})} \oplus \hat{x}_{2\ (\text{case 2})}$, $w_3 = \hat{x}_{4\ (\text{case 1})} \oplus \hat{x}_{4\ (\text{case 2})}$, and $w_4 = \hat{x}_{6\ (\text{case 1})} \oplus \hat{x}_{6\ (\text{case 2})}$ are utilized to discriminate the correct set of values between the case 1 and case 2 estimates, as presented hereafter.

- $\hat{x}_0, \hat{x}_2, \hat{x}_4, \hat{x}_6$ bits correction

In the error free scenario, $\hat{x}_{0\ (\text{case 1})}$ coincides with the value of $\hat{x}_{0\ (\text{case 2})}$ and thus $\{w_1, w_2, w_3, w_4\} = \{0, 0, 0, 0\}$.

If one error occurs in the sequence
$\left\{S^{(0)}, C_1^{(0)}, C_0^{(0)}, \ldots, S^{(3)}, C_1^{(3)}, C_0^{(3)}\right\}$, then one value of $\hat{x}_{(\text{case 1})}$ is computed wrong, and one of the $\{w_1, w_2, w_3, w_4\}$ bits is equal to "1". In this situation, the case 1 decoded bits $\{\hat{x}_0, \hat{x}_2, \hat{x}_4, \hat{x}_6\}$ are the correct ones.

If one error occurs in the sequence $\{E_1, E_2, E_3, E_4\}$, then three values of $\hat{x}_{(\text{case 2})}$ are computed wrong, and three of the bits $\{w_1, w_2, w_3, w_4\}$ are equal to "1". In this situation, the case 2 decode bits $\hat{x}_0, \hat{x}_2, \hat{x}_4, \hat{x}_6$ are the correct ones.

Thus, to summarize the discrimination bits are used as:

- If $w_1 + w_2 + w_3 + w_4 = 3$ then choose case 2 $\{\hat{x}_0, \hat{x}_2, \hat{x}_4, \hat{x}_6\}$ decoded bits;
- Otherwise, choose case 1 $\{\hat{x}_0, \hat{x}_2, \hat{x}_4, \hat{x}_6\}$ decoded bits.

- $\hat{x}_1, \hat{x}_3, \hat{x}_5, \hat{x}_7$ bits correction

Suppose one error occurred and the correct value of $\hat{x}_0$ was obtained using the above methodology. However, there are two Haar encoded 3-bit sequences for the correct value of $\hat{x}_0 = 0$, per se. Specifically, $(S^{(0)}, C_1^{(0)}, C_0^{(0)})$ is then either $(0,\ 0,\ 0)$ or $(0,\ 1,\ 1)$, which means we do not know exactly whether the value of $\hat{x}_1$ is "0" or "1". However, since bit E_5 is correct and equal to $C_1^0 \oplus C_1^{(1)} \oplus C_1^{(2)} \oplus C_1^{(3)}$, it allows for the determination of the correct value of $\hat{x}_1$. In this case $\hat{x}_1 = E_5 \oplus C_1^{(1)} \oplus C_1^{(2)} \oplus C_1^{(3)}$.

Note that the situation when one error affects E_5 is of no relevance, as all the 12 S and C bits required to restore the correct input byte $\{x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$ are correct.

Double Error Detection

A first case is when any two bits from the sequence $\left\{S^{(0)}, C_1^{(0)}, C_0^{(0)}, \dots, S^{(3)}, C_1^{(3)}, C_0^{(3)}, E_1, \dots, E_4\right\}$, are affected, which results in a value of $w_1 + w_2 + w_3 + w_4$ which is either equal to 2, or equal to 4.

A second case is when one bit-flip error is in the previous 16-bits sequence, while the other error affects bit E_5. In such a case, the error in the sequence can be detected with the previous single detection flow (otherwise stated, $w_1 + w_2 + w_3 + w_4$ has to be either equal to 1 to equal to 3), while bit E_5 can be duplicated and transmitted twice over the wires, at the extra cost of an additional redundant wire.

We note that even if full SECDEC protection requires up to 18 wires, if area is a foremost design optimization goal as in our case, a plausible alternative is to make use of both clock edges and transmit the 18 bits over a 9-wire bus (9 bits on the rising edge, and the other 9 bits on the falling edge).

6.3 Simulation Results

To gain insight on the practical implications of the proposed Haar codec enhanced data transport approach we evaluate by means of SPICE simulations the following systems:

- "Ref" - An 8-wire reference system, for which uncoded, raw data are transmitted over the wires. The 8-wire system serves as comparison reference from timing, energy, and area standpoints, as no coding scheme for optimizing the energy and reliability characteristics of the data transmission is applied in this case.

- "H12" (Haar) - A 12-wire system, which makes use of the scheme proposed in Section 6.2.
- "H8" - A 12-wire system which transmits encoded data by means of the scheme proposed in Section 6.2, the design being done with smaller distance between wires such that the bus width is preserved, i.e., $W(H8) = W(Ref)$.
- "BI" (Bus Invert) - A 9-wire system implementing the coding scheme in [96], a simple and efficient precursor of many coding methods and typical comparison reference.
- "Ref + Hamm" - A 12-wire system which corresponds to the "Ref" system protected with a single error detection and correction $(15, 11)$ Hamming code (8 bits information, 12 bits code length).
- "H12 + Hamm" - An 8-wire system which corresponds to the "H12" system protected with a single error detection and correction $(7, 4)$ Hamming code (4 bits information, 7 bits code length). The 3 Hamming bits protect the even bits $\{x_0, x_2, x_4, x_6\}$. One more parity bit protects the odd bits $\{x_1, x_3, x_5, x_7\}$. As in total $12 + 3 + 1 = 16$ bits are required to be sent over the bus, an 8-wire bus can be employed (the data being sent over the bus on both clock edges).
- "H12 + ECC" - A 9-wire system afferent to the "H12" system protected using the SECDED scheme proposed in Section 6.2. In this case a total of $12+5 = 17$ bits are required, which can be sent over a 9-wire bus with a double data rate (on both edges of the clock).

Figure 6.4 depicts the configuration to transmit the bits over the wires for each system. For each system, the SPICE simulation setup consists of encoders & input buffers, interconnect, and output buffers & decoders. Figure 6.3 depicts the simulation setup for the proposed interconnect codec augmented systems. The setup for the 8-wire, reference system is similar, with the exception of the encoder and decoder blocks which are excluded. As concerns the interconnect, for given specifications (e.g., wire length, number of parallel conductors), and technology parameters (e.g., related to the dielectric and metal layer stack conductivity, dielectric permittivity, wire pitch, aspect ratio etc.), a SPICE RLGC compatible model was obtained using the Synopsys Raphael electromagnetic field solver. The simulations were performed in SPICE, employing a commercial 45nm technology, at nominal operating conditions, for different bus lengths (spanning from 1mm to 10mm in increments of 1mm to cover medium and long range interconnects) and bus widths (spanning from 8 to 512, using multiple shielded 8-wire bus subsystems). As data to be transmitted over the wires, 10000 randomly generated bytes are provided as system input, one byte per clock cycle.

Subsequently all considered systems are analyzed with respect to area, delay, and energy interconnect performance monitors.

	Wire Number											
	w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	w11	w12
Ref	x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7				
H12; H8	$S^{(0)}$	$C_1^{(0)}$	$C_0^{(0)}$	$S^{(1)}$	$C_1^{(1)}$	$C_0^{(1)}$	$S^{(2)}$	$C_1^{(2)}$	$C_0^{(2)}$	$S^{(3)}$	$C_1^{(3)}$	$C_0^{(3)}$
Ref + Hamm	x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	H_1	H_2	H_2	H_3
H12 + Hamm	$S^{(0)}$	$C_1^{(0)}$	$C_0^{(0)}$	$S^{(1)}$	$C_1^{(1)}$	$C_0^{(1)}$	H_1	H_2				
	$S^{(2)}$	$C_1^{(2)}$	$C_0^{(2)}$	$S^{(3)}$	$C_1^{(3)}$	$C_0^{(3)}$	H_3	H_4				
H12 + ECC	$S^{(0)}$	$C_1^{(0)}$	$C_0^{(0)}$	$S^{(1)}$	$C_1^{(1)}$	$C_0^{(1)}$	E_1	E_2	E_5			
	$S^{(2)}$	$C_1^{(2)}$	$C_0^{(2)}$	$S^{(3)}$	$C_1^{(3)}$	$C_0^{(3)}$	E_3	E_4				

Figure 6.4: Bits transmission configuration.

6.3.1 Energy & Area

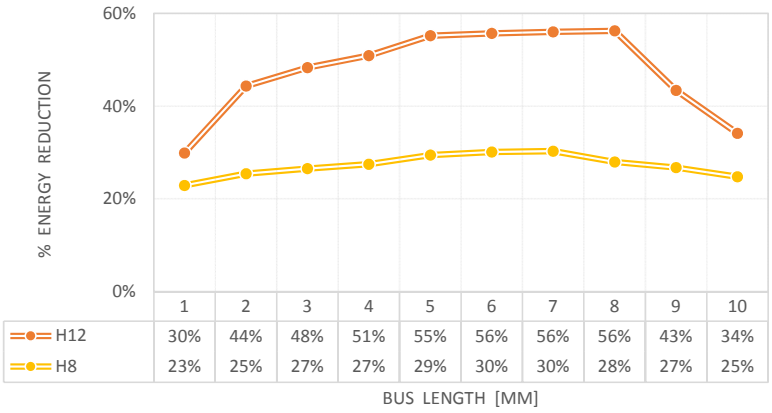
The consumed energy is measured for the entire system (encoder/input buffers + interconnect + decoder/output buffers), and over the entire duration of transmission (i.e., $10000 \times T_{clk}$). To provide a fair comparison, T_{clk} is tailored for each analyzed system (as a function of the wire length, and of the encoder/decoder maximum operation frequency), such that the data at each system output can be correctly sampled. The energy is measured in SPICE using the supply current integrated over the entire transmission duration

$$Energy = \int_0^{10000 \times T_{clk}} I_{VDD}(t) \cdot V_{VDD}\ dt,$$

thus we capture both the static and the dynamic energy components.

Energy Oriented Haar Systems

Figure 6.5 graphically illustrates the energy reduction achieved by the 8-bit "H12", "H8", and "BI" schemes for different interconnect lengths. The energy percentages are reported relative to the energy figures obtained for the Ref 8-wire system. A similar trend is observed for all systems: As the interconnect length increases, the energy gain also increases, which is as expected, since longer interconnects are more energy demanding than shorter ones, and thus can benefit more from a switching activity reduction on its wires. For the "H12" and "H8" systems, an energy gain apex of 56% and 30% is manifested at $6-8$mm and $6-7$mm, respectively. The lowest energy reduction potential is observed for shorter interconnects (30% and 23%, respectively for 1mm wires), where the codec energy is more significant, and thus the Haar scheme be-



	1	2	3	4	5	6	7	8	9	10
H12	30%	44%	48%	51%	55%	56%	56%	56%	43%	34%
H8	23%	25%	27%	27%	29%	30%	30%	28%	27%	25%

Figure 6.5: Energy profile vs. interconnect length.

comes less effective. When moving to longer, e.g., 10mm interconnects lower energy saving of 34% and 25% are achieved for the "H12" and "H8" systems, respectively, which can be attributed to the longer wires and driving buffers consumed energy counterpoising the codec enabled energy benefits. The "BI" design is clearly less effective than the Haar based designs and even results in energy increase for shorter than 5 mm wires.

As concerns the area footprint, a less than 1% hardware overhead corresponds to the Haar systems, which is not unexpected, as the codec requires very simple logic. Since the chip area footprint is determined by the logic and local interconnects, and not by the global interconnects, which are implemented in the upper metal layers, and their afferent vias, the 8 to 12 bus width increase does not add any additional overhead to the area footprint. Thus the "H12" system has $\approx$ the same area as the 8-wire "Ref", while from the energy point of view is more effective with 27% on average than the "Ref" system.

To assess the potential sensitivity of the codec augmented interconnect performance w.r.t. the bus width, we evaluated the energy reduction - as depicted in Figure 6.6 - for the transmission of a 512-bit wide input data vector over 8, 16, 32, 64, 128, 256, and 512-bit wide 5mm long buses (the Haar system bus being composed of multiple shielded 12-wire wide bus segments). We observe an energy gain inflexion point is obtained for the 64-wire bus (55%), w.r.t. which 6% and 2% energy increase is obtained for the 512-wire bus and 8-wire bus, respectively. We attribute this to: (i) the switching activity profile changes (e.g., as the bus size increases from S to $2S$ one potential bus switching is eliminated, thus leading to smaller energy dissipation), and (ii) the transmission time reduction (e.g., as the bus size increases from S to $2S$ the transmission time is reduced by half). On the other hand, as the bus size increases the parasitics also become more complex, and interconnect and area requirements are doubled, with negative implications on the timing and energy figures.

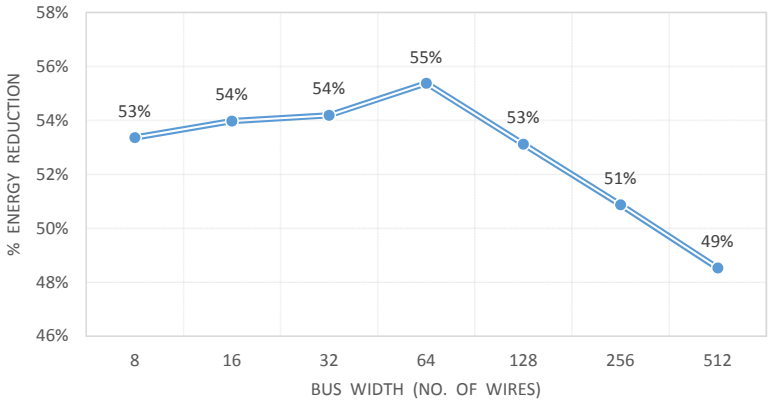


Figure 6.6: Energy profile for "H" system for bus length of 5mm vs. bus width.

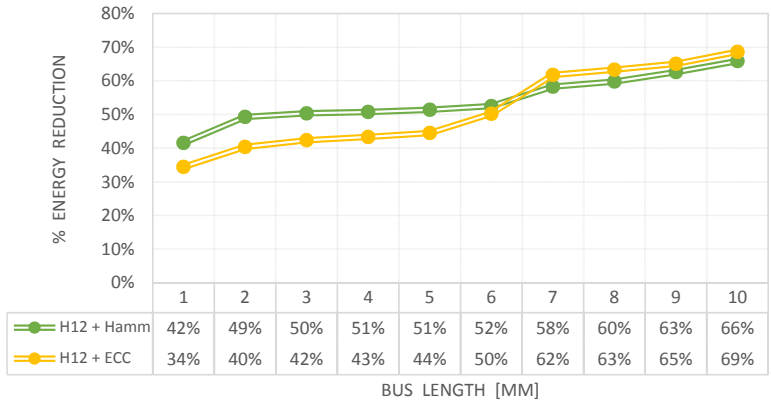
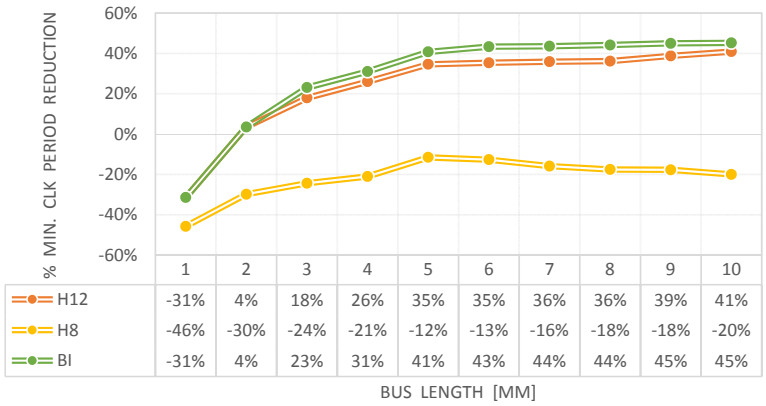


Figure 6.7: Energy profile vs. interconnect length for the ECC protected systems.

Energy and Reliability Oriented Haar Systems

Figure 6.7 depicts the percentage energy gain for 12-wire "Ref + Hamm", 8-wire "H12 + Hamm", and 9-wire "H12 + ECC" systems, relative to the energy figures of the 8-wire "Ref" system, for 1mm to 10mm bus lengths. We observe that the direct augmentation of the reference with SECDED capabilities (the "Ref + Hamm" system) comes with a great energy consumption increase (more than $2\times$ "Ref") with an aggravation trend as the interconnect length increases, which can be explained by a worse parasitics profile for the 12-wire bus and an activity profile higher in coupling transitions. Conversely the two Haar-based systems become more energy effective with the interconnect length increase and even consume less energy than the "Ref" unprotected baseline for longer than 7mm interconnects. For interconnect length $<$ 7mm, the two Haar schemes enhanced with error protection consume more energy



	1	2	3	4	5	6	7	8	9	10
H12	-31%	4%	18%	26%	35%	35%	36%	36%	39%	41%
H8	-46%	-30%	-24%	-21%	-12%	-13%	-16%	-18%	-18%	-20%
BI	-31%	4%	23%	31%	41%	43%	44%	44%	45%	45%

Figure 6.8: Minimum clock period vs. interconnect length.

than the ”Ref” system. Specifically, on average the 8-wire ”H12 + Hamm” and 9-wire ”H12 + ECC” consumes with 15% and 28% more energy than the ”Ref” system, respectively. However, for interconnect length above 7mm both ”H12 + Hamm” and ”H12 + ECC” are more energy effective than the ”Ref” system consuming 11% and 18% less energy on average, respectively. When compared to the 8-wire ”H12 + Hamm” system, the 9-wire ”H12 + ECC” systems enables an activity profile with fewer transitions on the extra error detection and correction lines, which is reflected in higher overall energy gain when the interconnect energy dissipation is the dominant contributor ($\geq$ 7mm). Note that the energy dissipated by the two ECC augmented Haar-based systems can be further diminished, if the power supply voltage is reduced (as the afferent timing faults single errors can be corrected by the SECDED logic). Area wise, the ECC augmented systems require 1%, $\approx$ 2%, and $\approx$ 2% area overhead with respect to the ”Ref” area, for ”Ref+Hamm”, ”H12 + Hamm”, and ”H12 + ECC”, respectively.

6.3.2 Delay

Energy Oriented Haar Systems

Simulation results reveal that for the ”Ref” 8-wire system, the bit arrival time for each wire exhibits a smaller spread when compared to the ”H12” system. However, the ”Ref” maximum arrival time is larger than the one provided by the Haar system counterpart, except for smaller length wires (1mm). This has positive implications on the transmission clock period, as indicated by Figure 6.8, which can be decreased with 35%, and 41% for 5mm, and 10mm, respectively. For 1mm, the clock period is negatively impacted, as it is increased with 31% vs. ”Ref”, which can be attributed to the effects of bus switching activity diminution that are more prominent for medium and longer wires than for shorter wires (the total delay - encoder+decoder+bus - of

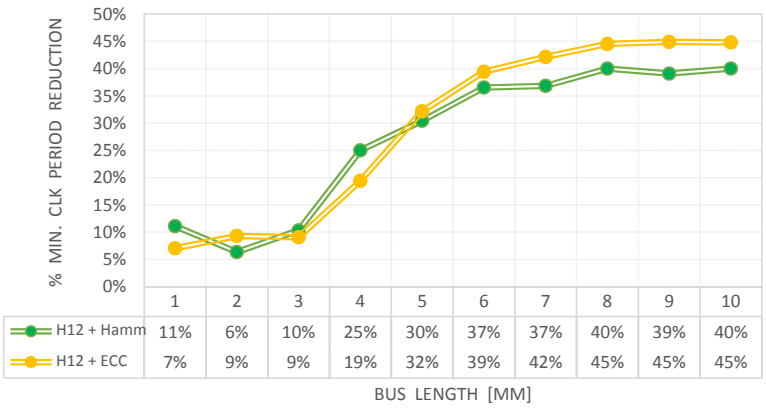


Figure 6.9: Minimum clock period vs. interconnect length for the ECC protected systems.

the coding-based system having to counterbalance the 8-wire bus delay, in order to obtain clock period benefits). The "H8" system is negatively impacted for all considered lengths, requiring a clock period increase vs. "Ref" of 46%, 12%, and 20% for 1mm, 5mm, and 10mm, respectively. We attribute this to the bus design parameters (e.g., decreased spacing), since the "H8" bus width is smaller than the one of the "H12" bus, occupying the same metal layer area as the "Ref" 8-wire bus. We note that the delay figures are afferent to the "H12" system without repeaters, reflecting the maximum wire length for which the signal integrity is preserved and which enables the maximum energy savings. If we buffer the "H12" system, the delay can be further improved, but at the expense of consuming extra energy, and increasing the area footprint. Additionally, the system without repeaters benefits the stringent time-to-market constraints, as it enables a faster timing-closure progress.

Energy and Reliability Oriented Haar Systems

Figure 6.9 depicts the percentage reduction of the minimum clock period for the ECC protected systems w.r.t. the "Ref" clock period. The "Ref + Hamm" system clock period is decreased for all wire lengths, while for the "H12 + Hamm" and "H12 + ECC" systems, the frequency is decreased only until 7mm and 6mm, respectively. At 10mm the "H12 + Hamm" system can properly operate at a clock period 9% smaller than "Ref" clock period, while the "H12 + ECC" system enables a clock period decrease of 16% w.r.t. "Ref". We note that even if for the two Haar systems, the propagation delay across the wires is reduced, the added delay of the ECC/Hamming encoder and decoder makes the overall delay reduction gain possible only for longer wires. We note that the reported delay figures correspond the the minimum bus delay for which the signal integrity is preserved. However, the delay can be further reduced under the safe operation value, since an ECC scheme is in place and can correct potential errors.

6.4 State-of-the-Art Comparison

Subsequently, we give a brief account of the most recent prevalent articles documenting bus coding, and a comparison against the state-of-the-art performance figures for the "H12" and the "H12 + ECC" systems. We note that a direct comparison with state-of-the-art is not always straightforward, e.g., when the implementations are done in different technologies, to which effect we apply the Dennard scaling [110], in the case of an analytical evaluation, or one which doesn't account for the codec performance penalties.

- *Power Reduction Codes*

Generally speaking the power-reduction oriented bus coding research corpus can be broadly divided into two main categories: methods which reduce the per wire self-switching activity and methods which reduce both self and coupling (between neighboring wires) transitions. The former category disregards the inter-wire coupling parasitics, which for VDSM technologies results in delay and power penalties. The latter category is in better alignment with state of the art interconnect physical phenomena and as a result enables power savings. However, delay and reliability aspects are not considered. With a few exceptions (e.g., Bus Invert coding [96], a simple and efficient precursor of several methods), the preponderant existing low power bus coding techniques are only effective either for data buses (e.g., [100–102]) or for address buses (e.g., nonredundant permutation-based codes - Gray code [103], data dependent reordering codes [104]; redundant codes - Odd/Even Bus Invert [105] which extends [96] for coupling activity reduction, $T0$ code [106], Beach code [107], Limited Weight code [108])). Furthermore, most of these techniques mainly exploit the spatial and/or temporal correlations of the transferred data, which render them less effective for random data transmission. Following the same philosophy, most recent work includes Conditionally Coded Blocks (CCB) code, Sign Extension (SEM) code, XOR/XNOR code, and Qaudro code [95].

State-of-affairs CCB and SEM [95] yield 58% and 60%, respectively power reduction in 130nm for a 16-bit transmission. The Quadro coding [95] achieves up to 47% reduction for byte-wise transmission, the technology node not being specified. Comparatively, we obtain in 45nm 56% energy reduction for 6mm to 8mm long 8-bit buses, and 54% energy reduction for 5mm 16-bit buses (2 shielded 8-bit bus segments).

- *Delay Reduction Codes*

Crosstalk Avoidance Codes (CACs) were proposed to reduce crosstalk induced delay by forbidding certain transitions (e.g., opposite direction switching) on adjacent wires [97] or bit patterns (e.g., $'010'$ and $'101'$) [98], which would cause the highest delay (afferent to the worst adjacent capacitive coupling cases). However, as integral part of an IC, interconnects are exposed to various environmental aggression factors (e.g., supply voltage fluctuation, electromagnetic interference), which pose in the absence of an error resiliency mechanism, signal integrity and reliability problems.

[98] reports 21% power savings in a 90nm process with an encoder area of 369 2-input gates for 12-bit transmission, while [97] estimate a delay of 300ps in 65nm for a

pipelined implementation, at the expense of 17% less area overhead than [98]. In our case the operating frequency can be increased by 35% for 5mm, with a lightweight hardware implementation (a codec gate count of 16 and total logic depth of 3 vs. a state-of-art gate count in the order of hundreds), thus surpassing state-of-the-art.

- *Reliability Improvement Codes*

To combat errors that may occur during bus transmission and thereby diminish the necessity and impact on performance of timing and voltage margins over-designing traditional approach, error control codes were explored, most notably Hamming and cyclic linear block codes [99]. However, they usually incur significant performance penalties caused by the codecs high complexity, which is not the case for the proposed SECDED Haar based systems.

- *Joint Codes*

Combining in a serial manner low power codes with crosstalk avoidance codes proved to be inefficient, as the crosstalk avoidance properties per se, will be altered/canceled by a subsequent low power coding. However, crosstalk avoidance codes benefit also, as an aside the power consumption (even if to a lesser extent when compared to the savings achieved by low power codes). Conversely, error control codes can be combined with either low power codes or crosstalk avoidance codes, towards joint power/delay-reliability merits, with the chief caveat of significant coding overhead, as the joint code is merely a concatenation of two individual, independent codes. In this regard, very few joint coding techniques have been investigated, e.g., [111, 112], which attain single and double, respectively error correcting capacity, by combining a Hamming code with crosstalk avoidance codes; single error correcting [113] and [114] which introduce the Duplicate-Add-Parity code, Modified-Dual-Rail code and Boundary-Shit code, respectively. State-of-affair ECC based methods [113] report $\approx$ 40% energy reduction when compared to standard Hamming code for a byte-wise transmission, using 10mm wires in 130nm, while our savings w.r.t. standard Hamming are 1.2$\times$ bigger in 45nm. [114] achieves the ECC capability for 9 extra wires and 45 FO4 codec gate delay, while [111] reports a codec delay of 27 FO4 (unpipelined) and an energy dissipation 15.1% and 18.3% lower over [113] and [114], in 130nm.

To conclude, our overall simulation results indicate significant energy savings, while increasing the operating frequency, and having also SECDED capabilities, thus outperforming state-of-the-art counterparts.

6.5 Conclusion

In this chapter we proposed an energy effective bus coding scheme, that facilitates a higher operating frequency when compared to the uncoded counterpart, as well as with other equivalent state-of-art counterpats. We further augmented the Haar codec with a tailored SECDED scheme with energy, area, and frequency merits outperforming a direct ECC augmentation of the uncoded system. We analyzed two systems: the Haar-based system which targets energy efficiency, and the SECDED augmented Haar-

based system which targets a reliable and energy effective data transport. Simulation results in 45nm of the Haar system for an 8-wire interconnect and various workload profiles, indicate energy savings of 55% and 34% and operation frequency increase of 35% and 41% for 5mm and 10mm, respectively, at the expense of an area increase of less than 1%, with respect to the reference uncoded system. When compared to the reference system (not SECDED protected), for 10mm long buses, the Haar tailored SECDED approach consumes 27% less energy at the expense of 2% area overhead, while providing robuster data transmission.

Given that in order to create a reliable computation platform one needs the means for reliable data transport, storage, and processing, and state of the art reliable data storage (memories) rely on ECCs we investigate in the next chapter the potential such an approach may have in the implementation of reliable functional (data processing) units out of unreliable devices.

Note. The content of this chapter is based on the following paper:

N. Cucu Laurenciu, Y. Wang, S. D. Cotofana, **Haar-based Interconnect Coding for Medium/Long Range Energy Effective, Reliable Data Transport**, submitted to Design, Automation and Test in Europe (DATE), (2017).

7

Reliable Data Processing Units

The significant uncertainty associated with current nanodevices fabrication and operation, calls for a circuit design paradigm change, which ought to actively embrace the inherently nanodevice unreliability to generate overall circuit architectures able to perform reliable computation. While for storage units coding theory based viable solutions exist, Data Processing Units (DPUs) are not amenable to a similar line of reasoning. The typical approach undertaken for fault-tolerant DPUs relies on modular redundancy (e.g., spatial, temporal), which while being effective from an error tolerance perspective, generally involves high area and/or performance impairments. This chapter proposes a generic methodology to obtain reliable DPU implementations built with unreliable components by intimately intertwining Error-Correcting Codes (ECCs) codecs with the DPU functionality. The ECC protected DPU architecture is derived cluster-wise with area and reliability constraints, by exploiting dependence relations (logical and w.r.t. area shared) between internal signals pertaining to the DPU and ECC codec. To evaluate the error rate and performance implications, a multitude of test corners were considered (e.g., gate criticality, ECC type and structure, faulty and low complexity decoder, time-space redundancy) for an ECC protected 6-bit adder architecture. Simulation results reveal that the ECC embedding approach can be effective from both error rate and area perspective, for the Pareto designs with performance figures of merit situated in-between consecutive modular redundancy based design corresponding curves. The proposed approach is generic from the coding point of view, scalable, and enables a fine grained control of the DPU desired reliability degree and area overhead.

7.1 Introduction

As CMOS technology down scaling is approaching fundamental limitations (e.g., photolitographic manufacturing, physical as the transistor feature reaches the size of a few atoms, and economical with diminishing returns) the nanodevices physical and electrical parameters variability is greatly increased [81]. As a result, the devices no longer function deterministically but, their de facto operation is rather governed by a

stochastic process, which in turn significantly impacts their reliability. Furthermore, even when operating correctly, the nanodevices exhibit an increased susceptibility to transient and permanent errors. On the other hand, nascent technologies, e.g., single electron tunneling, carbon nanotubes, molecular devices, while holding promise as alternatives to silicon semiconductors, are still in their infancy and deficitary in mass production means [115]. Therefore, a seemingly natural avenue for the CMOS technology scaling is to perform reliable nanoscale computing with unreliable devices, which reiterates once again the early vacuum tubes challenge.

To this effect, in the case of nanoscale memories, Error-Correcting Codes (ECCs) are traditionally used, proving to be a viable solution [36], [37]. The data are encoded with an ECC prior to their storage and afterwards transient bit flipping faults can be detected and corrected periodically. For digital logic however a similar line of reasoning is not applicable any longer, as in this case it is not the ECC protected memory content that is directly affected by errors, but the hardware (the logic gates) whose correlated and cumulated errors effect is reflected in the DPU ECC protected outputs. Thus while for memories ECC redundancy is generated only as a function of the data to be protected, prior to its storage, for DPUs ECC redundancy has to be generated during the computation of the data to be protected. The conventional solution for logic is to employ some form of von Neumann style modular redundancy, which can be implemented in either time (e.g., repeat the same computation on the same hardware), or space (e.g., execute the same computation in parallel on hardware replicas) [22]. From an information theoretic point of view, modular redundancy (in both space and time) rely on repetition coding. For modular redundancy the output is expected to be the same even if it undergoes different computation conditions (i.e., (i) for temporal redundancy different environmental conditions, and (ii) for spatial redundancy different hardware replicas, which may be implemented in different technologies, logic styles etc.). This is similar to transmitting the same data over different channels. While benefiting from simplicity of implementation, the modular redundancy approach is rather inefficient from the coding point of view. Thus arises the natural question of whether an ECC with a better error correction capability which requires only a small constant overhead in order to protect a circuit logic against random bit-flip faults can be employed.

The main research question that we address in this chapter can be formulated as follows: Given a fabrication technology able to provide basic circuit components, i.e., logic gates, with an error rate of 10^{-x} in certain environmental conditions, can we systematically derive a circuit topology able to implement a given Boolean function F such that the circuit output Word Error Rate (WER) is 10^{-y} with $y > x$ (noting that y can be significantly smaller than x if no fault tolerance technique is used), while abiding to a set of design constraints in terms of area, delay, and power consumption? To this end, we propose to augment the traditional algebra design avenue with information theory concepts, i.e., to make use of ECCs in order to improve the reliability of circuits built out of unreliable devices. The key element behind this approach is the realization of an intimate intertwining between the original circuit whose Primary Outputs (POs) are desired to be protected via ECC, and the afferent codec hardware.

Specifically, the ECC protected circuit (whose outputs may or may not include the original circuit POs, depending on the ECC type) is synthesized in independent clusters based on certain criteria (i.e., logical dependencies and amount of area shared) while targeting a Word Error Rate (WER) after decoding that is smaller than the ECC protected circuit and the original circuit gate error rate. The detailed methodology for deriving the ECC protected circuit architecture is presented in Section 7.2. We employed for simulation a 6-bit Brent-Kung parallel prefix adder and investigated the influence of several codes and design strategies upon the protected adder WER/area merits. As described in Section 7.3, we evaluated a multitude of test corners by: (i) varying ECC parameters (error correction capacity) and structure (ECC type, modularized ECC), (ii) varying gate criticality (w.r.t. circuit POs inflicted damage) to identify an optimum WER/area design point, and (iii) considering different architectural optimizations for both the encoder and decoder (gate criticality aware synthesis for the encoder, low complexity architecture for the decoder). Simulation results reveal that the proposed approach can be effective from both WER and area perspective for the Pareto designs with performance figures of merit situated in-between consecutive modular redundancy based design counterparts. Section 7.4 concludes the chapter with a summary of the main points and results.

7.2 General Framework and Simulation Setup

Subsequently, a systematic methodology for implementing fault tolerant data processing units by means of ECCs is described. In a nutshell, our problem, graphically conceptualized in Figure 7.1, can be defined as: Given a combinational logic circuit subjected to faults inducing conditions, augment the original circuit with an ECC codec able to protect the circuit POs while being itself subjected to errors, such that after decoding, the correct, error-free original circuit outputs can be recovered. In Figure 7.1, the protected circuit POs form a codeword of the embedded ECC, which, in the error free case, is the same as the one obtained by encoding the output of F. Hence, the logical functionality of the ECC protected circuit is the same as the serial concatenation of F with the ECC encoder, but its hardware implementation is derived as a function of F PIs (note that in practice it doesn't make sense to encode the output of F once it has been computed, since in this case the ECC decoder will attempt to recover the input of the ECC encoder, i.e., the possibly erroneous POs computed by F). The crux is the ability to intimately intertwine the ECC codec and the original circuit, enabling a fault tolerant Boolean function synthesis. More precisely, based on the circuit topology, dependencies subject to certain constraints (e.g., reliability, area) between the encoded outputs (which may or may not include the original circuit POs depending on the used code) are identified and used for driving the logic synthesis process of the ECC protected circuit. For evaluation, we selected based on decoding complexity and code performance rationale the following block linear codes [116]: (i) Reed-Müller codes, and (ii) linear codes optimal w.r.t. the code length, for given error correction capacity and information size. To assess the merits of the designs

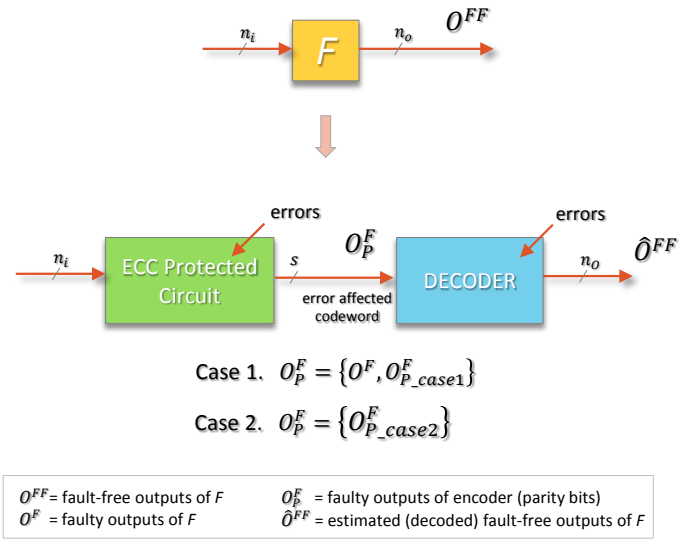


Figure 7.1: ECC Protection for Boolean logic F.

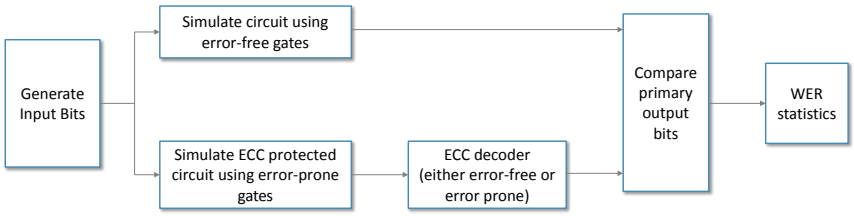


Figure 7.2: Reliability Evaluation Framework.

enhanced with error protection in terms of area and reliability, we consider the following two metrics: (i) total number of 2-input gates, as preliminary area indicator, and (ii) WER as reliability monitor, the word under observation being composed of the original circuit POs. As concerns the area evaluation, the logic synthesis is performed using Cadence RTL Compiler [55] for a commercial 40nm *LP* CMOS technology node. For a fair comparison, only 2-input basic gates were employed. The reliability evaluation flow follows the block scheme depicted in Figure 7.2. A statistically relevant batch of 10 million circuit input vectors were generated via a maximum period 31-bit shift register pseudo-random number generator. The original circuit without any codec, assumed error-free in this case (no errors injected), is then simulated and the obtained PO values are saved to provide the reference for subsequently deriving the WER statistics. For the ECC codec enhanced circuit reliability simulation, we consider homogeneous (identical) gate probability of failure. We concern ourselves with transient and permanent faults manifested for logic gates as single bit-flips of

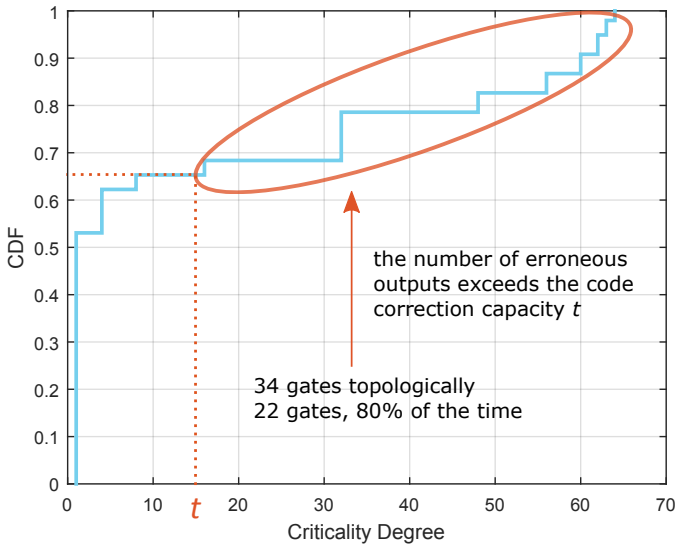


Figure 7.3: Gates Criticality Statistics.

their outputs. For bit-flip inflicting on a gate output with a given probability of failure, the *xorshift*128+ generator [117] is employed. Once the original, error-free circuit and the ECC codec enhanced error-prone one are both simulated, their POs are compared to each other in order to derive the WER statistics of interest. We selected a logarithmic delay Brent-Kung parallel prefix adder as discussion vehicle, as it is highly utilized in state of the art processor implementations. A 6-bit architecture is considered to maintain the simulation time within reasonable limits and thus allow us to investigate a large number of codes and design strategies. Let assume an ECC with length $n = 64$, dimension $k = 7$ (sum, carry out), and error correction capacity $t = 15$. We note that a circuit gate will have a different impact on the WER, as each faulty gate can possibly affect a certain maximum number of POs depending on the circuit topology. To quantify the faulty gates impact on the circuit WER, we introduce the gate criticality notion defined as the number of POs that one can reach starting from the output of that gate. Figure 7.3 illustrates the Cumulative Distribution Function (CDF) of the criticality degree for all gates pertaining to the ECC protected adder. We note that $\approx 65\%$ of the total number of gates have a criticality less than the code correction capacity t (thus the ECC protected adder POs can be corrected when any of these gates is faulty), while the remaining 35% of the gates, topologically speaking, can induce more than t errors to the ECC protected circuit outputs (and thus diminish the ECC efficiency). Furthermore, multiple gates may be faulty concomitantly, even if individually each faulty gate has a criticality $< t$. Hence it is desirable for gates to exhibit a criticality as small as possible w.r.t. the threshold t. On the other hand, one may note that the gate criticality is merely a topological upper bound, as in practice, due to fault masking and preponderant workload patterns, the maximum number of outputs

that can actually be affected by a faulty gate is less. Indeed, when subjecting the ECC protected adder to a random 10000 13-bit random workload, of a total of 34 gates with a topological (maximal) criticality $> t$, only 22 gates actually affect 80% of the time more than t outputs. While this suggests a reduced number of gates (22) targeted for circuit redesign in view of criticality reduction below threshold t (e.g., via gate replication), it still remains a prohibitively complex process to impose constraints on individual gates, while maintaining the error resiliency characteristic of the remaining circuit structure, and at the same time keeping the overall area footprint within reasonable limits. In view of this, a more judicious architectural solution would be to consider a higher level of granularity, i.e., a gate clustering approach. Specifically, to derive the ECC protected circuit architecture, we propose the following methodology: **Step 1.** The ECC protected circuit POs (i.e., the codec encoded bits which are a function of the original circuit POs, and thus also of the original circuit PIs) are first aggregated in groups of x, with x being the desired maximum gate criticality. The aggregation criteria are given by the outputs affinity w.r.t. the area shared between them. **Step 2.** RTL synthesis is then performed with area/timing/power constraints for each group of POs. Thus each group of POs is synthesized as a function of all the original circuit PIs, and has its own cone of logic, independent of the other groups cones. In this way, within each cone of logic, the reliability constraints (i.e., the gate criticality $\leq$ the number of group outputs), are always satisfied.

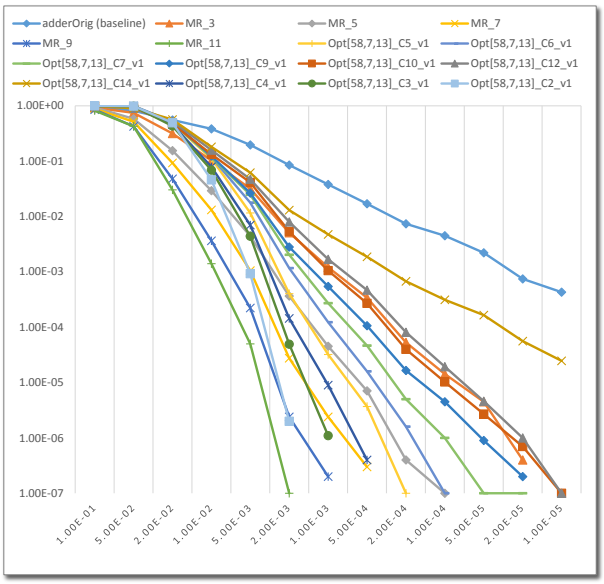
7.3 Design Space Exploration

Having presented the methodology to derive the ECC protected circuit architecture, we are now in position to investigate the implications of various architectural and ECC related strategies, from WER and area standpoints. Subsequently, we adhere to the notation MR-z for a modular redundancy scheme and ECC$[n, k, t]$_Cx_vy, where (i) ECC can be RM for Reed-Müller codes, or Opt for optimal codes; (ii) $[n, k, t]$ denote the ECC parameters - code length (encoded bits) n, info size (bits to be protected) k, and error correction capacity t; (iii) Cx denotes the maximum gate criticality; and (iv) vy is used for architecture versioning.

7.3.1 Gate Criticality

To explore the gate criticality influence on WER, we consider an optimal systematic code Opt$[58, 7, 13]$ and vary the ECC protected adder gate criticality from 2 to 14.

Figure 7.4 illustrates the WER and area overhead of the ECC protected adder relative to the baseline design and the MR protected counterparts. The figure reveals that the area overhead is increasing as the gate criticality decreases. This is obvious since the outputs are divided in more groups and as a result the amount of area shared gets smaller. Conversely the WER curves improve with the gate criticality decrease. The reason is twofold: (i) there is a smaller probability that faulty gates will corrupt multiple outputs (as there is less area shared between the group outputs),



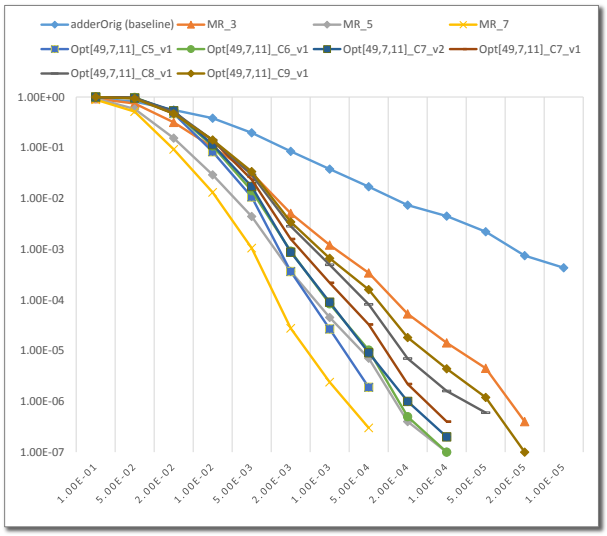
Design	# gates	× increase
AdderOrig (baseline)	**65**	
Opt[58,7,13]_C2_v1	835	12.8
Opt[58,7,13]_C3_v1	613	9.4
Opt[58,7,13]_C4_v1	500	7.8
Opt[58,7,13]_C5_v1	415	6.4
Opt[58,7,13]_C6_v1	361	5.6
Opt[58,7,13]_C7_v1	329	5.1
Opt[58,7,13]_C9_v1	283	4.3
Opt[58,7,13]_C10_v1	247	3.8
Opt[58,7,13]_C12_v1	220	3.4
Opt[58,7,13]_C14_v1	209	3.2
MR_3	195	3.0
MR_5	325	5.0
MR_7	455	7.0
MR_9	585	9.0
MR_11	715	11.0

Figure 7.4: WER/Area vs. Gates Criticality.

and (ii) there is a higher probability of fault masking occurrence (up to a certain design inflexion point after which the bigger area facilitates less and less fault masking). Summarizing, when varying the gate criticality, efficient design points can be identified between the MR curves, with effective WER/area trade-offs. For instance the design Opt$[58, 7, 13]_C9_v1$ has an area penalty of $4.3\times$ baseline (which is 14% less than the MR-5 area and 43% more than the MR-3 area) and a WER curve that lies approximately in the middle between the MR-3 and the MR-5 WER curves.

7.3.2 Code Correction Capacity

For this test corner, 8 codes optimal w.r.t. the codeword length n for fixed information size k (i.e., 7) and an error correction capacity t spanning from 5 to 17 were investigated ($[n, k, t] \in \{ [26, 7, 5], [31, 7, 6], [34, 7, 7], [42, 7, 9], [49, 7, 11], [58, 7, 13], [63, 7, 15], [74, 7, 17] \}$). Figure 7.5 illustrates the WER and area figures obtained for the Opt$[49, 7, 11]$ code. One can observe: (i) the design Opt$[49; 7; 11]$_C7_v1 exhibits a WER that is very close to the MR-5 one, while area wise it requires 10% less area, (ii) the design Opt$[49; 7; 11]$_C8_v1 has a WER curve that is situated approximately in the middle between the MR-3 and MR-5 curves, with an area overhead of 33% relative to the MR-3 area, but with 20% less area when compared to the MR-5 area, and (iii) the design Opt$[49; 7; 11]$_C9_v1 shows a WER curve that is slightly better than the MR-3 one, while requiring 33% more area, trade-off which is not justified as in the (i) and (ii) cases. In the figure we plotted for the Opt$[49; 7; 11]$ code, the



Design	# gates	× increase
AdderOrig (baseline)	**65**	
Opt[49,7,11]_C5_v1	387	6.0
Opt[49,7,11]_C6_v1	359	5.5
Opt[49,7,11]_C7_v2	330	5.1
Opt[49,7,11]_C7_v1	291	4.5
Opt[49,7,11]_C8_v1	262	4.0
Opt[49,7,11]_C9_v1	259	4.0
MR_3	195	3.0
MR_5	325	5.0
MR_7	455	7.0

Figure 7.5: WER/Area for Code Correction Capacity $t = 11$.

merits of only a few designs which correspond to a gate criticality of 5 to 9 and certain clusterization optimization points. However, several design architectures were obtained with WER in-between two consecutive MR WER curves (e.g., between the MR-3 and MR-5 WER curves). Thus, from the previous results it can be concluded that the coding approach offers a much higher flexibility and achievable WER range granularity when compared to the MR approach. Even though for the design points with WER curves approaching the MR ones, the area penalty is higher which renders the coding approach rather inefficient, in-between the MR curves, the ECC approach enables WER curves and area figures which are bounded by two consecutive modular redundancy figures. For instance between the MR-3 and MR-5 WER curves, several intermediary curves can be obtained via ECC, at the expense of an area whose value is lower bounded by the MR-3 area and upper bounded by the MR-5 area. This indicate that, as a function of the gate criticality degree and the check bits clustering policy efficiency, such an ECC embedding design with good performance (WER as well as area) relative to two consecutive modular redundancy designs, can always be found.

7.3.3 ECC Type

Both systematic and non-systematic block linear codes were considered. The systematic code has the advantage that the original circuit POs are part of the codeword. This implies that the check bits clusterization can either exclude those bits identical with the original circuit POs (in which case the original circuit architecture is preserved as it is in a single cluster - encoding non-intrusiveness which is most often desirable) or include all codeword check bits. For the non-systematic code,

only the latter situation is applicable as the original circuit POs are not part of the codeword but, are obtained indirectly via decoding. For evaluation we considered Reed-Müller codes as the same decoder architecture can be employed for both the systematic and non-systematic forms. It was observed that the systematic RM code behaves better than the non-systematic counterpart from both the WER and area standpoints. Next we considered optimal length codes for fixed error capacity and information size [118] [119]. When compared to RM codes with similar codeword length, the RM codes enabled better WER and area figures. However, when the code correction capacity and its length are varied, the optimal codes provide a slightly better alternative to shortening and puncturing of the Reed-Müller codes. From the point of view of the decoder architectural complexity, the Reed-Müller ECCs have a significant advantage over the optimal codes, as they can be decoded efficiently using majority logic. Summarizing, as simulation results indicate, systematic Reed-Müller codes seem to present the greatest potential among the linear block codes, with respect to the codec architectural complexity, as well as with induced WER figures.

7.3.4 Time-Space Redundancy

As a functional unit (e.g., an adder) workload is usually not sporadic in time, but rather consists in batches of multiple consecutive operations, we investigated whether a spatial redundancy approach could be taken advantage of, en route to WER improvement. Specifically, a single code could be employed to protect not only the result of one addition, but the results of p consecutive additions altogether. We continue the discussion subsequently assuming a systematic code, however a similar rationale holds true for non-systematic codes. A straight forward approach is a spatial-only approach, i.e., to make use of p adder replicas, as illustrated in Figure 7.7, each with its own different set of inputs. The codeword consists of (i) $p \times n_o$ bits provided by each adder replica F and (ii) the remaining m parity bits which are derived by the subcircuit P as a function of all p adders primary inputs. This perspective incurs however a high area penalty, which is not justified by its corresponding WER improvement. To alleviate the area overhead, a compound spatio-temporal redundancy is sought. Concretely, this translates to: (i) instead of replicating p times the adder, a single instance is used and subjected to p consecutive workloads to generate the first $p \times n_o$ codeword bits, and (ii) the remaining m codeword bits are computed as in the previous spatial-only redundancy approach, as a function of all the p adders PI values. For practical evaluation, we first explored a time redundancy factor of 4, which translates to 28 bits to be protected via ECC (RM$[128, 28, 15]$). We obtained a WER curve which lies between the MR-3 and MR-5 WER curves, while area-wise is $31.8\times$ baseline, as compared to $3\times$ and $5\times$ baseline for the MR-3 and MR-5 designs, respectively. The high area is a direct consequence of the high codeword size (128), as the encoded bits are a function of all 8 6-bit adder operands and 4 carry-in bits. A relatively high criticality degree (7) was employed to diminish the area overhead but its further increase would result in a WER worse than the MR-3 one. In an effort to reduce the number of ECC bits, and thus the total area footprint,

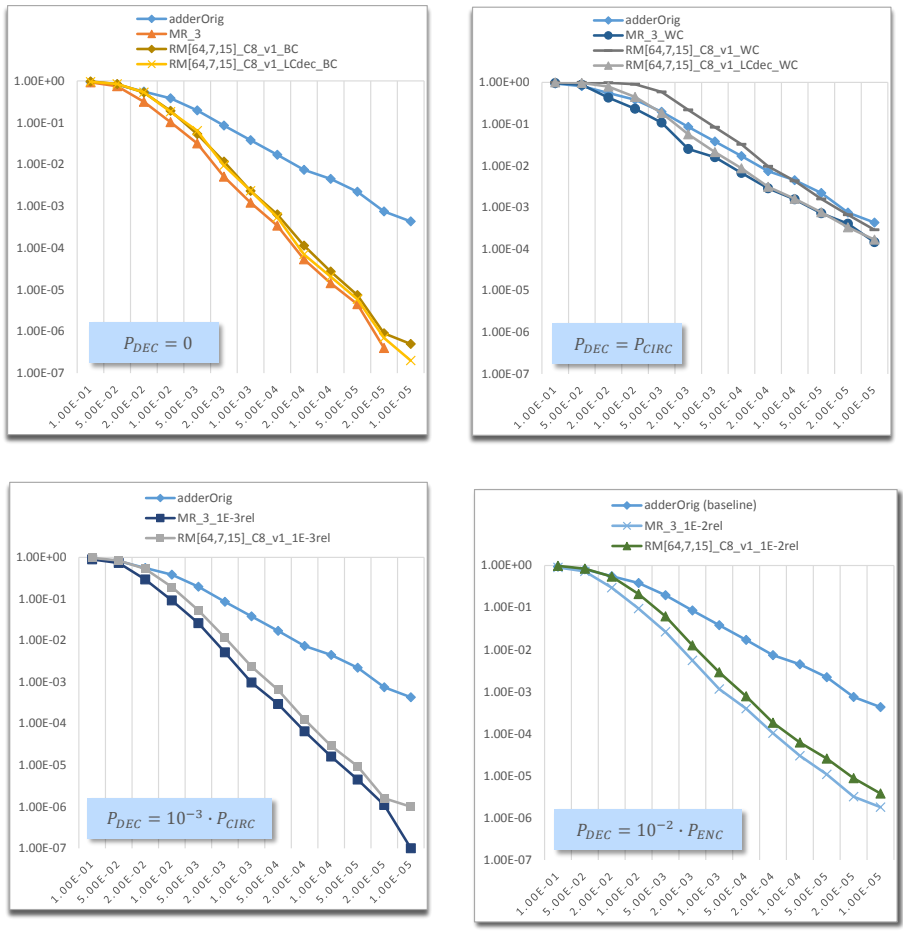


Figure 7.6: WER for Non-Ideal Decoding.

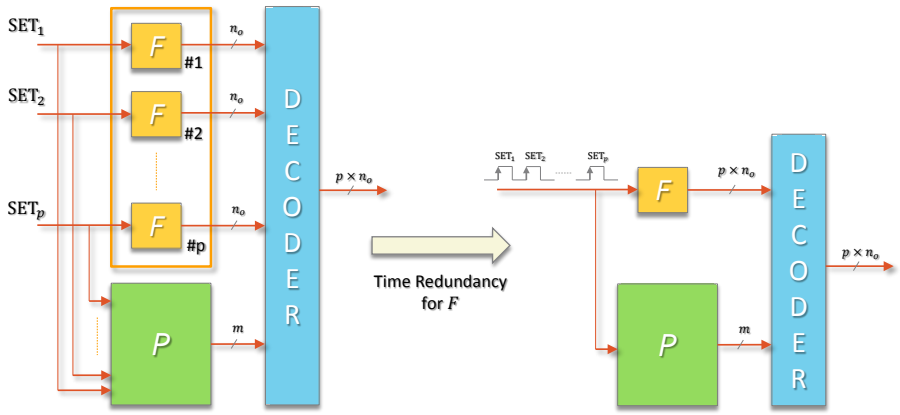


Figure 7.7: Time-Space Redundancy.

we considered next a time redundancy factor of 2, which translates to 14 bits that necessitate ECC protection, resulting in a 55-bit codeword ((RM[55, 14, 7]) code). As expected, the area decreased substantially by more than 3× w.r.t. the previous analyzed cases, to yield 9.7× baseline. While the WER curve is slightly better than the MR-3, this comes at a still high area penalty increase (i.e., ≈ 3 × MR-3).

7.3.5 Non-Ideal Decoding

In all experiments thus far we assumed an ideal error-free decoding. However, this assumption may induce a significant WER bias as the decoder is subjected to the same environmental conditions as the ECC protected circuit and original circuit. Generally speaking, the decoder can be designed with tighter margins and reliability constraints, such that in-field it exhibits a much lower error rate when compared to the ECC protected circuit or the original circuit. In view of this, we considered different non-systematic Reed-Müller codes, and evaluated the WER of the system composed of ECC protected circuit and decoder. The decoder gate failure probability is set to a fraction of the one affecting the ECC augmented adder circuitry. Otherwise stated, two gate failure probability islands are considered inside the codec, which realistically can be realized by over-design or by making use of two fabrication technology nodes, i.e., an advanced one for the ECC protected adder and an older and more reliable one for the decoder. For simulations, a one-step parameterizable Reed-Müller majority logic decoder was employed. To grasp the WER bias an ideal decoder can induce we plotted in Figure 7.6 the WER curves for ideal vs. faulty Reed-Müller decoder, for different decoders gate probability of failure: (i) WC - worst case, i.e., the decoder gates probability of failure is equal to the ECC protected circuit gates probability of failure, (ii) BC - best case, i.e., the decoder is error-free, (iii) 1E-3rel - the decoder gates probability of failure is $10^3\times$ less than the ECC protected circuit counterpart, and (iv) 1E-2rel - the decoder gates probability of failure is $10^2\times$ less than the ECC

protected circuit counterpart. As depicted in Figure 7.6 the MR-3 and RM designs exhibit similar WER curves (slightly better for the MR-3). When compared to the original adder WER, Figure 7.6 reveals slightly better WER curves for the MR-3 and RM designs, when both the ECC protected circuit and decoder operate under the same gate failure probability. The 1E-3rel case gives almost identical WER as the BC case, while the WER curves for the 1E-2rel case lie approximately in the middle of the BC and WC case. As intuited, the decoder probability of failure has a big impact on the system WER, and hence shouldn't be disregarded. From the area overhead standpoint, the Reed-Müller decoder becomes the dominant contributor to the system area, increasing it from $4.9\times$ baseline (software ideal decoding) to $41.7\times$ baseline (hardware decoding). This substantial increase is attributed to a big extent to the decoding logic afferent to the first decoded bit, whose computation complexity is much higher than the one associated to all the rest decoded bits.

To alleviate this problem, the first bit can be treated in a different manner, i.e., by means of MR-3. As first bit can be utilized the functional unit PO bit whose triplication induces the least area penalty, which in the adder case equates to the LSB sum bit. This translates to a decoder implementation of 32 XOR2 gates followed by a majority voter for each of the 6 MSB bits, and a majority 3 voter for the LSB bit. A total area reduction of $\approx 50\%$ is obtained by the previously described technique when compared to the original Reed-Müller decoder case. WER-wise improvement is observed for both the BC and WC cases - as illustrated in Figure 7.6 by the curves with extension LCdec - when compared to the original Reed-Müller decoder BC and WC cases.

A second improvement of the Reed-Müller decoder is related to the manner voting ties are treated in the majority logic. As the number of parity relations for decoding each of the adder POs is even, typically an extra input wired to either logic "1" or logic "0" is utilized for each majority voter to achieve a relevant (odd) voting. However, in this manner systematic errors are introduced. A more appropriate alternative is to use a random number generator as the extra input, but the associated area cost prohibits its use. In view of these, a different path is pursued, by noting that extra adder PO bit values can come from the ECC protected circuit clusters at no extra area penalty (as they are already internal signals part of the ECC protected circuit clusters logic - the encoded bits logical equations are expressed as functions of the original adder POs, which implies that for each cluster, all adder POs involved in the encoded bits computation are inherently determined). We propose to exploit this fact, by using the clusters internal signals which correspond to the adder POs as extra inputs to the decoder majority voting logic, to enable an unbiased voting with practically no area penalty incurred. Figure 7.8 illustrates the architecture of the low complexity decoder for a 3-bit adder.

7.3.6 Modularized ECC

Encouraged by the error protection efficient trends for storage, we investigated whether a similar line of reasoning can be applied in the context of functional units.

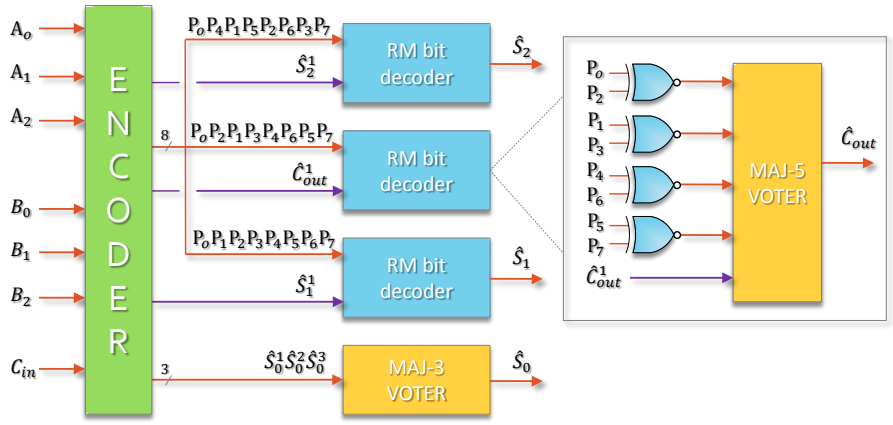


Figure 7.8: Reed-Müller Low Complexity Decoder for 3-Bit Adder.

While for memories the error correction code is applied directly on the bits to be protected, in the case of functional units, the code protects the results of a circuit logic upon which the code has no scope of influence. Thus, for the same error probability, the word error rate at a memory storage output is much smaller than the word error rate at a circuit POs. The reason behind this is that the memory content is directly affected by errors, while in the case of logic circuits, the comprising gates are affected by errors, and it is their cumulated and correlated effect that is reflected at the primary outputs. While for memories a SECDED Hamming code is typically the state of the art approach, for functional units it is far from being effective, because of the much higher error rate experienced by the information bits. We already seen in Section 7.3.4 that WER efficiency is not gained by increasing the number of information bits to be protected, and implicitly by using a longer length ECC, as it is the case for memories. Nevertheless, this result is conducive to the reverse phenomenon, i.e., employing instead of a single long length ECC multiple smaller length ECCs, each protecting its own share of the functional unit POs. To this end we split the adder in two parts and consider two Reed-Müller codes: RM$[26, 3, 3]$ for the carry out and the 2 sum MSBs, and RM$[25, 4, 3]$ for the 4 sum LSBs. We obtained a WER curve that is very close to the MR-3 WER, with an area overhead of $4.1\times$ baseline, which is only 27% bigger than the MR-3 $3\times$ baseline area. However, for less area requirements $3.9\times$ the baseline, the WER figures afferent to the single longer ECC based design are considerably improved when compared to the designs with two modularized ECCs, which renders the modularized ECC endeavour rather inefficient.

7.4 Conclusion

In this chapter we investigated whether the avenue of functional units reliability improvement via intimately ECC embedding could prove efficient. We propose a generic methodology to derive the ECC protected DPU architecture cluster-wise with area and reliability constraints, by exploiting dependence relations (logical and w.r.t. shared area) between internal signals pertaining to the DPU and the ECC codec. An extensive design space exploration was performed by considering a multitude of architectural and ECC related test corners (gate criticality, ECC type and structure, faulty and low complexity decoder, time-space redundancy). Simulation results for a 6-bit adder reveal the ECC protection approach can be effective from both Word Error Rate and area perspective for the Pareto designs with performance merits situated between consecutive modular redundancy figures. For instance, the adder protected with an optimal code w.r.t. code length and with an error correction capacity of 5, has an area overhead of 33% relative to the 3-modular redundancy design, but with 20% less than the 5-modular redundancy design, while from the error rate point of view it is situated approximately in the middle of the error rate characteristics of the two modular redundant designs. Additionally our approach eliminates the modular redundancy Achilles' heel as the centralised majority voting is decentralized in our case, which potentially results in robuster circuit implementations. The proposed approach is generic from the coding point of view, scalable and enables a fine grained control of functional units desired reliability degree.

Note. The content of this chapter is based on the following paper:

N. Cucu Laurenciu, T. Gupta, V. Savin, S. D. Cotofana, **Error Correction Code Protected Data Processing Units**, Proceeding of the 12^{th} ACM/IEEE International Symposium on Nanoscale Architectures (NANOARCH 2016), (2016).

8

Conclusions and Future Work

In this thesis, we have presented an integrated reliability aware design and life-time management framework, and pursued 3 main investigation avenues: (i) circuit reliability assessment and End-of-Life prediction for aging mitigation/compensation purpose, (ii) aging sensors design and raw sensor data processing towards circuit level aging information, and (iii) reliable data transport and processing via unreliable devices. As concerns (i), we first introduced a transistor-level aging model, which takes into account both the intrinsic device degradation and the degradation inflicted by neighbor devices. Subsequently, we proposed a circuit-level aging model, which assesses circuit level reliability status based on the aging status of a relevant set of comprising transistors. Addressing (ii), we first proposed a sensor design which can detect the amalgamated effect on a circuit block of concomitantly active failure mechanisms. Based on the noisy aging sensor measurements, we subsequently proposed a model able to predict a circuit End-Of-Life (EOL) at run-time. We followed (iii) by proposing: (a) codec assisted data transport structures, able to deal with technology scaling related effects, and (b) Error Correcting Codes (ECCs) protected Data Processing Units (DPUs).

8.1 Summary

The thesis contents can be summarized as follows:

Chapter 1 - In Chapter 1, we discussed the main ICs reliability issues of current technology nodes, and argued the necessity and relevance of an integrated reliability-aware framework from design-time to run-time, in order to design and enable dependable computing with ICs made from unreliable components. We introduced such a holistic framework, formulated the corresponding research questions addressed in the thesis, and highlighted the proposed approaches.

Chapter 2 - In Chapter 2 we introduced a transistor level aging model, in the context of obtaining reliability optimized circuit designs. While certain low-level parameters

may reflect better a certain aging mechanism effect, their unification into a general model which accounts for the multiple, concomitant degradation mechanisms effect is a very complex task. To this end, we employ as aging quantifier, a high-level transistor parameter, i.e., the output voltage slope at half the supply voltage, which is merely a functional of the low-level aging parameters as functions of time. The model accounts for the transistor intrinsic self-degradation (via own low-level parameters, e.g., threshold voltage), as well as for the influence of adjacent degraded transistors (via the gate voltage slope and the source current slope). To this end, we define the transistor age as the time integral of the sum of output voltage slope variations caused by the aging induced drift of intrinsic and extrinsic parameters. Our experimental results performed on a CMOS 2-input NOR gate indicate that: (i) the transistor output voltage slope proves to be an accurate aging measure able to capture the combined result of own degradation due to working regime and environmental conditions and of other devices ages which influence the correct functioning of the transistor under study, and (ii) aging is underestimated when the surrounding circuit topology influence is not taken into account.

Chapter 3 - In Chapter 3 we introduced two methods to assess the reliability status at circuit-level based on the reliability status of a reduced size set of transistors that are the most relevant from the aging standpoint (i.e., those which contribute the most to the overall circuit aging). Specifically, the first circuit aging model predicts the IC EOL as the sum of the critical transistors EOL values, weighted by fixed topology-dependent coefficients and it is more suited for run-time estimations. The second is a Markovian model accounting for the joint effects of process, environmental, and temporal variations. It is more more appropriate for design-time estimates and allows the circuit performance boundary to vary in time such that both remnant and non-remnant variations are encompassed for a better fitting to real conditions, and more appropriate guard-bands selection and aging mitigation/compensation techniques. The proposed transistor selection methodology applied on the ISCAS-85 $c499$ circuit implemented in a 45nm CMOS technology provided a critical transistors kernel of 15 from the total of 1526 transistors. Using this kernel and the first method, a small approximation error for the circuit estimated EOL was obtained (mean $< 15\%$ and standard deviation $< 6\%$, when compared with Cadence simulation results. We also demonstrated that further trade-offs area vs. EOL estimation accuracy can be done by reducing the transistor kernel cardinality to e.g., 5, case in which an approximation error with mean $< 10\%$ and standard deviation $< 15\%$ was obtained. When applying the statistical Markovian framework, we obtained a circuit aging approximation error with mean $< 10\%$ and standard deviation $< 15\%$ relative to the values obtained with Cadence. As the Markovian model takes into consideration more parameters and aging sources, these estimated end-of-life values which are further from the Cadence counterparts may be closer to the real end-of-life values but for the time being we do not have the means to validate this conjecture.

Chapter 4 - In Chapter 4, we proposed an on-chip aging sensor able to directly detect the joint effect of multiple degradation mechanisms, by measuring the peak power

supply current of a circuit block or entire IC. We note that unlike existing sensors, which monitor the aging of individual transistors, the proposed sensor can monitor in principle the entire circuit which means thus that no transistor kernel identification and models for aging abstractization from transistor to circuit level are required in our case. Simulations performed for a set of ISCAS−85 benchmark circuits indicate that the sensor reports a similar aging rate as the one indicated by Cadence simulation for most of circuit lifetime. However, towards circuit EOL, the peak power supply current and the threshold voltage aging indications start to deviate from each other. Unfortunately, no other aging reference is available thus we cannot tell which method provides better estimates for that circuit lifetime part. We note however that high accuracy is essential during an IC early life, when it undergoes the greatest wearout induced degradation, demanding thus accurate and fast reliability assessment and EOL prediction for an appropriate degradation compensation/mitigation measure.

Chapter 5 discusses the run-time circuit EOL estimation based the sensor aging measurements. As sensor measurements are uncertain to a certain extent (because of noise and measurements imperfections) and the circuit degradation is nonlinear, we propose to model the circuit degradation as a Wiener process with nonlinear mean, which is expressed as a combination of basis functions, weighted by degradation history dependent parameters. Based on the entire degradation history, and not only on the instantaneous degradation state, the degradation model parameters are updated via Bayesian inference, once new degradation data are accumulated. Simulation results indicate that the proposed framework may be better suited for capturing the nonlinear degradation dynamics when compared to the commonly employed models with linear mean. This is of relevance especially for the circuit early life stages, when the extent of degradation is the greatest, and as such the degradation combating techniques should be most effective.

Chapter 6 - In Chapter 6, we introduced and evaluated Haar codec assisted data transport via medium and long range unreliable interconnect targeting energy savings at the expense of a reasonably small area overhead, i.e., 1 extra wire, a 2-gate encoder, and a 2-gate decoder, for each and every pair of uncoded wires. As evaluation reference we considered an 8-bit synchronous bus transmission. We evaluated the proposed codec-assisted transport scheme for a 45nm CMOS technology, and obtained energy savings of (55% and 34% for 5mm and 10mm, respectively, and transmission frequency increase of (35% and 41% for 5mm and 10mm, respectively, at the expense of less than 1% area overhead with respect to the reference system (i.e., 8-wire synchronous uncoded bus). Later on, we augmented the codec with a Single Error Detection and Correction scheme that is particularly suited for the codec architecture. When compared to the reference system (not SECDED protected), for 10mm long buses, the Haar tailored SECDED approach consumes 27% less energy at the expense of 2% area overhead, while providing robuster data transmission.

Chapter 7 - In Chapter 7, we proposed a generic methodology to obtain reliable data processing unit implementations built with unreliable components by intimately inter-

twining Error-Correcting Codes (ECCs) codecs with the computing unit functionality. While viable ECCs solutions have been devised for unreliable storage units, computing units are not amenable to a similar line of reasoning, as in this case it is not the ECC protected memory content that is directly affected by errors, but the hardware (the logic gates) whose correlated and cumulated errors effect is reflected in the DPU ECC protected outputs. The error rate and performance implications of the proposed approach were evaluated for a multitude of corners (e.g., gate criticality, ECC type and structure, faulty and low complexity decoder, time-space redundancy) for an ECC protected 6-bit adder architecture, and revealed that the ECC embedding approach can be effective from both error rate and area perspectives, for the Pareto designs with the performance figures located in-between consecutive modular redundancy based designs corresponding curves. For instance, the adder protected with an optimal code w.r.t. code length and with an error correction capacity of 5, has an area overhead of 33% relative to the modular redundancy 3 design, but with 20% less than the modular redundancy 5 design, while from the error rate point of view it is situated approximately in the middle of the error rate characteristics of the two modular redundant designs. Furthermore, the proposed approach eliminates the single point of failure typical for the modular redundancy approaches (the majority voter).

8.2 Future Research Directions

Subsequently, we outline future research directions suggested as a continuation of the research avenues undertaken in this thesis.

1 – Statistical CAD for Joint Area-Delay-Power-Reliability Design Space Exploration
In this thesis, in the context of the design time sub-framework depicted in Figure 1.4, we have addressed the aging assessment at the transistor/circuit level related issues (i.e., models and the degradation sensing infrastructure). The aging models are further utilized at design time to guide the 4-dimensional (area, delay, power, reliability) design exploration process within an appropriate Circuit Assisted Design (CAD) tool, in order to obtain a reliability optimized circuit realization. Most of the CAD tools rely on deterministic foundations and do not take reliability into account. Thus a CAD tool able to perform constrained multi-objective optimization, and potentially migrate towards statistical algorithms is desired.

2 – Design of Aging Mitigation/Compensation Mechanisms
Another component that we have not addressed in this thesis is constituted by the aging mitigation/compensation mechanisms to be included in the design time sub-framework reliability wrapper, as illustrated in Figure 1.4. Several schemes can be designed, their selection for being appended to the reliability optimized circuit realization depending on their aging alleviation effectiveness under different stress profiles for the reliability aware circuit realization.

3 – Run-Time Reliability Management Policy Algorithms
From the run time sub-framework, illustrated in Figure 1.5, we concerned ourselves in this thesis with the aging assessment, EOL prediction and WER evaluation related issues, which are further used to guide the reasoning about adopting particular reliability management strategy if the assessed circuit reliability is not compliant with the QoS specifications. Algorithms for policy selection are thus desired in this context.

Bibliography

[1] J. E. Lilienfeld, "Method and Apparatus for Controlling Electric Currents," in *U. S. Patent No. 1,745,175*, 1926. (Cited on page 1.)

[2] O. Heil, "Improvements in or Relating to Electrical Amplifiers and Other Control Arrangements and Devices," in *British Patent No. 439, 457*, 1935. (Cited on page 1.)

[3] J. Bardeen and W. Brattain, "Three-Electrode Circuit Element Utilizing Semiconductor Materials," in *U. S. Patent No. 2,524,035*, 1948. (Cited on page 1.)

[4] J. A. Hoerni, "Method of Manufacturing Semiconductor Devices," in *U. S. Patent No. 3,025,589*, 1959. (Cited on page 1.)

[5] R. N. Noyce, "Semiconductor Device-and-Lead Structure," in *U. S. Patent No. 2981877*, 1959. (Cited on page 1.)

[6] "Computer History Museum," 2015. `http://www.computerhistory.org`. (Cited on page 1.)

[7] A. Masaki, Y. Harada, and T. Chiba, "200-Gate ECL Master-Slice LSI," in *IEEE International Solid-State Circuits Conference*, 1974. (Cited on page 1.)

[8] R. Beeson and H. Ruegg, "New Forms of All-Transistor Logic," in *IEEE International Solid-State Circuits Conference*, 1962. (Cited on page 1.)

[9] "The Story of the Intel 4004," 2016. `http://www.intel.com/content/www/us/en/history/museum-story-of-intel-4004.html`. (Cited on page 2.)

[10] G. Moore, "Cramming More Components onto Integrated Circuits," in *Electronics*, vol. 38, pp. 114–117, 1965. (Cited on page 2.)

[11] R. H. Dennard and et al., "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions," in *IEEE Journal of Solid State Circuits (JSSC)*, vol. SC-9, pp. 256–268, 1974. (Cited on page 2.)

[12] "GARTNER INC.," 2015. `http://www.gartner.com/technology/core/products/research/markets/semiconductorIndustryDevices.jsp`. (Cited on page 2.)

[13] M. Pecht, *Product Reliability, Maintainability and Supportability Handbook.* 2nd edition, CRC Press, 2009. (Cited on page 3.)

[14] M. Ohring, *Reliability and Failure of Electronic Materials and Devices.* Academic Press, Elsevier, 1998. (Cited on page 4.)

[15] J. H. Lala and R. E. Harper, "Architectural Principles for Safety-Critical Real-Time Applications," in *Proceedings of the IEEE*, vol. 82, pp. 25–40, 1994. (Cited on page 4.)

[16] "The Star Computer.," 2016. `http://history.nasa.gov/computers/Ch5-5.html`. (Cited on page 4.)

[17] "The Saturn Guidance Computer.," 2016. `http://history.nasa.gov/computers/Ch7-2.html`. (Cited on page 4.)

[18] "Software Implemented Fault Tolerance System.," 2016. `https://www.sri.com/work/timeline-innovation/`. (Cited on page 4.)

[19] J. H. Lala, "A Byzantine Resilient Fault Tolerant Computer for Nuclear Power Plant Applications," in *16th Annual International Symposium on Fault Tolerant Computing Systems (FCTS)*, vol. 18, pp. 338–343, 1986. (Cited on page 4.)

[20] "Number 1 Electronic Switching System.," 2016. https://en.wikipedia.org/wiki/Number_One_Electronic_Switching_System. (Cited on page 4.)

[21] S. Grandhi and et al., "Reliability Aware Logic Synthesis through Rewriting," in *27th IEEE International System-on-Chip Conference (SOCC)*, pp. 274–279, 2014. (Cited on page 9.)

[22] C. Huang, *Robust Computing with Nano-Scale Devices: Progresses and Challenges.* Springer Netherlands, 2010. (Cited on pages 9, 12, and 100.)

[23] N. Aymerich, S. D. Cotofana, and A. Rubio, "Controlled Degradation Stochastic Resonance in Adaptive Averaging Cell-Based Architectures," in *IEEE Transactions on Nanotechnology*, vol. 12, pp. 888–896, 2013. (Cited on pages 9 and 12.)

[24] M. A. Alam, "A Critical Examination of the Mechanisms of Dynamic NBTI for PMOSFETs," in *IEEE International Electron Devices Meeting (IEDM)*, 2003. (Cited on page 10.)

[25] M. A. Alam and et al., "A Comprehensive Model for PMOS NBTI degradation: Recent Progress," in *Microelectronics Reliability*, vol. 47, pp. 853–862, 2007. (Cited on page 10.)

[26] M. Ershov and et al., "Dynamic Recovery of Negative Bias Temperature Instability in p-type MetalOxideSemiconductor Field-Effect Transistors," in *Applied Physics Letters*, vol. 83, p. 1647, 2003. (Cited on page 10.)

[27] A. W. Strong and et al., *Reliability Wearout Mechanisms in Advanced CMOS Technologies.* Wiley-IEEE Press, 2009. (Cited on pages 10, 31, and 77.)

[28] W. Wang and et al., "Statistical Prediction of Circuit Aging under Process Variations," in *IEEE Custom Integrated Circuits Conference (CICC)*, pp. 13–16, 2008. (Cited on pages 10, 24, and 36.)

[29] B. Vaidyanathan, X. Oates, A. S.and Yuan, and Y. Wang, "NBTI-Aware Statistical Circuit Delay Assessment," in *Quality of Electronic Design (ISQED)*, pp. 13–18, 2009. (Cited on pages 10, 24, and 36.)

[30] Y. Lu and et al., "Statistical Reliability Analysis Under Process Variation and Aging Effects," in *46th ACM/IEEE Design Automation Conference (DAC)*, pp. 514–519, 2009. (Cited on pages 10, 24, and 36.)

[31] S. Han and J. Kim, "NBTI-Aware Statistical Timing Analysis Framework," in *IEEE International SOC Conference (SOCC)*, pp. 158–163, 2010. (Cited on pages 10 and 24.)

[32] S. Kumar, C. H. Kim, and S. Sapatnekar, "An Analytical Model for Negative Bias Temperature Instability," in *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pp. 493–496, 2006. (Cited on pages 11 and 36.)

[33] S. Bhardwaj and et al., "Predictive Modeling of the NBTI Effect for Reliable Design," in *IEEE Custom Integrated Circuits Conference (CICC)*, pp. 189–192, 2006. (Cited on pages 11 and 36.)

[34] X. Li, "Asymptotic Probability Extraction for Non-Normal Distribution of Circuit," in *IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, pp. 2–9, 2004. (Cited on pages 11 and 36.)

[35] S. Kumar, J. Li, C. Talarico, and J. Wang, "A Probabilistic Collocation Method based Statistical Gate Delay Model Considering Process Variations and Multiple Input Switching," in *Design, Automation and Test in Europe (DATE)*, vol. 2, pp. 770–775, 2005. (Cited on pages 11 and 36.)

[36] H. Naeimi and A. DeHon, "Fault Tolerant Nano-Memory with Fault Secure Encoder and Decoder," in *2nd International Conference on Nano-Networks*, pp. 1–7, 2007. (Cited on pages 12 and 100.)

[37] S. Ghosh and P. D. Lincoln, "Dynamic Low-Density Parity Check Codes for Fault-Tolerant Nano-Scale Memory," in *Foundations of Nanoscience*, 2007. (Cited on pages 12 and 100.)

[38] S. Borkar, "Role of Interconnects in the Future of Computing," in *Journal of Lightwave Technology*, pp. 3927–3933, 2013. (Cited on pages 12 and 82.)

[39] T. H. Kim, R. Persaud, and C. H. Kim, "Silicon Odometer: An On-Chip Reliability Monitor for Measuring Frequency Degradation of Digital Circuits.," in *IEEE Symposium on VLSI Circuits*, pp. 122–123, 2007. (Cited on pages 13 and 54.)

[40] J. Keane, X. Wang, D. Persaud, and C. Kim, "An All-In-One Silicon Odometer for Separately Monitoring HCI, BTI and TDDB.," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 4, pp. 817–829, 2010. (Cited on pages 13 and 54.)

[41] E. Karl, P. Singh, D. Blaauw, and D. Sylvester, "Compact In-Situ Sensors for Monitoring Negative-Bias-Temperature-Instability Effect and Oxide Degradation," in *IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 410–423, 2008. (Cited on pages 13, 15, and 54.)

[42] M. Agarwal, B. C. Paul, Z. Ming, and S. Mitra, "Circuit Failure Prediction and Its Application to Transistor Aging," in *25th IEEE VLSI Test Symposium*, pp. 277–286, 2007. (Cited on pages 13, 15, and 54.)

[43] M. Agarwal and et al., "Optimized Circuit Failure Prediction for Aging: Practicality and Promise.," in *IEEE International Test Conference*, pp. 1–10, 2008. (Cited on pages 13, 15, and 54.)

[44] S. Feng and et al., *Maestro: Orchestrating Lifetime Reliability in Chip Multiprocessors*, pp. 563–568. 2010. (Cited on page 14.)

[45] M. G. Moghaddam, A. Yamamoto, and C. Ababei, "Investigation of DFVS Based Dyanmic Relaibility Management for Chip Multiprocessors," in *International Conference on High Performance Computing and Simulation (HPCS)*, pp. 563–568, 2015. (Cited on page 14.)

[46] W. Song, S. Mukhopadhyay, and S. Yalamanchili, "Architectural Reliability: Lifetime Reliability Characterization and Management of Many-Core Processors," in *IEEE Computer Architecture Letters*, vol. 14, pp. 103–106, 2015. (Cited on page 14.)

[47] W. Wang, V. Balakrishnan, B. Yang, and Y. Cao, "Statistical Prediction of NBTI-Induced Circuit Aging," in *9th International Conference on Solid-State and Integrated-Circuit Technology (ICSICT)*, pp. 416–419, 2008. (Cited on page 23.)

[48] B. Paul, K. Kang, H. Kufluoglu, M. Alam, and K. Roy, "Impact of NBTI on the Temporal Performance Degradation of Digital Circuits," in *IEEE Electron Device Letters*, vol. 26, pp. 560–562, 2005. (Cited on page 23.)

[49] Y. Wang and et al., "Temperature-Aware NBTI Modeling and the Impact of Input Vector Control on Performance Degradation," in *Design, Automation and Test in Europe (DATE)*, pp. 1–6, 2007. (Cited on page 23.)

[50] T. Sakurai and A. R. Newton, "Delay Analysis of Series-Connected MOSFET Circuits," in *IEEE Journal of Solid State Circuits*, vol. 26, pp. 122–131, 1991. (Cited on page 23.)

[51] D. Lorenz, M. Barke, and U. Schlichtmann, "Efficient Analyzing the Impact of Aging Effects on Large Integrated Circuits," in *Microelectronics Reliability*, vol. 52, pp. 3–8, 2012. (Cited on page 24.)

[52] H. Luo and et al., "Circuit-Level Delay Modeling Considering both TDDB and NBTI," in *12th IEEE International Symposium on Quality Electronic Design (ISQED)*, pp. 1–8, 2011. (Cited on page 24.)

[53] D. Lorenz, G. Georgakos, and U. Schlichtmann, "Aging Analysis of Circuit Timing Considering NBTI and HCI," in *15th IEEE International On-Line Testing Symposium (IOLTS)*, pp. 1546–1552, 2009. (Cited on page 24.)

[54] E. Consoli, G. Giustolisi, and G. Palumbo, "An Accurate Ultra-Compact I-V Model for Nanometer MOS Transistors With Applications on Digital Circuits," in *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 59, pp. 159–169, 2011. (Cited on pages 26, 27, and 30.)

[55] "Cadence," 2012. `https://www.cadence.com/en/default.aspx`. (Cited on pages 30, 46, 63, 65, 77, and 102.)

[56] "BSIM Compact Model for Common Multi-Gate (MG) FETs," 2012. `http://www-device.eecs.berkeley.edu/~bsim/`. (Cited on page 31.)

[57] T. Siddiqua, S. Gurumurthi, and M. R. Stan, "Modeling and analyzing NBTI in the presence of Process Variation," in *12th International Symposium on Quality Electronic Design (ISQED)*, pp. 1–8, 2011. (Cited on page 36.)

[58] Y. Wang and S. D. Cotofana, "A Novel Virtual Age Reliability Model for Time-to-Failure Prediction," in *IEEE International Integrated Reliability Workshop Final Report (IRW)*, pp. 102–105, 2010. (Cited on pages 36 and 70.)

[59] K. R. Castleman, *Digital Image Processing.* Prentice Hall, 1995. (Cited on pages 37 and 43.)

[60] S. Sapatnekar, *Timing.* Springer, 2010. (Cited on page 38.)

[61] L. Breiman and J. H. Friedman, "Predicting Multivariate Responses in Multiple Linear Regression," in *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 59, pp. 3–54, 1997. (Cited on page 39.)

[62] B. E. Barrett and J. B. Gray, "A Computational Framework for Variable Selection in Multivariate Regression," in *Statistics and Computing*, vol. 4, pp. 203–212, 1994. (Cited on page 39.)

[63] M. Yuan and Y. Lin, "Model Selection and Estimation in Regression with Grouped Variables," in *Journal of the Royal Statistical Society: Series B*, vol. 68, pp. 49–67, 2006. (Cited on page 40.)

[64] J. A. Tropp, "Algorithms for Simultaneous Sparse Approximation, Part II: Convex Relaxation," in *Signal Processing*, vol. 86, pp. 589–602, 2006. (Cited on page 40.)

[65] H. Liu, M. Palatucci, and J. Zhang, "Blockwise Coordinate Descent Procedures for the Multi-Task Lasso, with Applications to Neural Semantic Basis Discovery," in *26th Annual International Conference on Machine Learning*, pp. 649–656, 2009. (Cited on page 40.)

[66] T. Simila and J. Tikka, "Input Selection and Shrinkage in Multiresponse Linear Regression," in *Computational Statistics and Data Analysis*, vol. 52, pp. 406–422, 2007. (Cited on page 40.)

[67] B. A. Turlach, W. N. Venables, and S. J. Wright, "Simultaneous Variable Selection," in *Technometrics*, vol. 47, pp. 349–363, 2005. (Cited on page 40.)

[68] K. S. Eshbaugh, "Generation of Correlated Parameters for Statistical Circuit Simulation," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 11, pp. 1198–1206, 1992. (Cited on page 42.)

[69] M. Dietrich and J. Haase, *Process Variations and Probabilistic Integrated Circuit Design.* Springer, 2012. (Cited on page 42.)

[70] X. Pan and H. Graeb, "Lifetime Yield Optimization of Analog Circuits Considering Process Variations and Parameter Degradations," in *Advances in Analog Circuits*, ch. 6, pp. 130–146, InTech, 2011. (Cited on page 43.)

[71] J. Medhi, *Stochastic Processes.* New Age Science, 2009. (Cited on page 44.)

[72] "BSIMProPlus," 2005. `http://www.proplussolutions.com`. (Cited on pages 46 and 77.)

[73] "Predictive Technology Model (PTM).," 2013. `http://ptm.asu.edu/`. (Cited on pages 46 and 77.)

[74] "Synopsys," 2012. `http://www.synopsys.com/home.aspx`. (Cited on page 46.)

[75] J. E. Chung, P. K. Ko, and C. Hu, "A Model for Hot-Electron-Induced MOSFET Linear-Current Degradation Based on Mobility Reduction Due to Interface-State Generation," *IEEE Transactions on Electron Devices*, vol. 38, no. 6, pp. 1362–1370, 1991. (Cited on page 58.)

[76] L. Xiaojun, Q. Jin, and J. B. Bernstein, "Compact Modeling of MOSFET Wearout Mechanisms for Circuit-Reliability Simulation," *IEEE Transactions on Device and Materials Reliability*, vol. 8, no. 1, pp. 98–121, 2008. (Cited on page 58.)

[77] G. O. Ducoudray, R. Gonzalez-Carvajal, and J. Ramirez-Angulo, "A High-Speed Dynamic Current Sensor for IDD test Based on the Flipped Voltage Follower," in *Southwest Symposium on Mixed-Signal Design*, pp. 208–211, 2003. (Cited on page 59.)

[78] Toumazou, C. and Hughes, J. B. and Battersby, N. C., *Switched-Currents: An Analogue Technique for Digital Technology.* The Institution of Engineering and Technology, 1993. (Cited on page 59.)

[79] E. Sackinger and W. Guggenbuhl, "A High-Swing, High-Impedance MOS Cascode Circuit," in *IEEE Journal of Solid-State Circuits*, vol. 25, pp. 289–298, 1990. (Cited on page 62.)

[80] G. Kim, M. K. Kim, B. S. Chang, and W. Kim, "A Low-Voltage, Low-Power CMOS Delay Element," in *IEEE Journal of Solid-State Circuits*, vol. 31, pp. 966–971, 1996. (Cited on page 62.)

[81] "International Technology Roadmap for Semiconductors (ITRS)," 2013. `http://www.itrs.net`. (Cited on pages 69 and 99.)

[82] J. Y. Zhao, F. Liu, and Q. Sun, "On-line Reliability Estimation and Performance Prediction for Metalized Film Pulse Capacitor," in *Acta Armamentarii*, vol. 27, pp. 265–268, 2006. (Cited on pages 69 and 75.)

[83] C. Y. Peng and S. T. Tseng, "Mis-Specification Analyses of Linear Degradation Models," in *IEEE Transactions on Reliability*, vol. 58, pp. 444–455, 2009. (Cited on pages 69 and 75.)

[84] C. C. Tsai, S. T. Tseng, and N. Balakrishnan, "Mis-Specification Analyses of Gamma and Wiener Degradation Processes," in *Journal of Statistical Planning and Inference*, vol. 141, pp. 3725–3735, 2011. (Cited on pages 69 and 75.)

[85] W. Wang, M. Carr, W. J. Xu, and A. K. Kobbacy, "A Model for Residual Life Prediction Based on Brownian Motion With an Adaptive Drift," in *Microelectronics Reliability*, vol. 51, pp. 285–293, 2010. (Cited on page 70.)

[86] X. S. Si and et al., "Remaining Useful Life Estimation Based on a Nonlinear Diffusion Degradation Process," in *IEEE Transactions on Reliability*, vol. 61, pp. 50–67, 2012. (Cited on page 70.)

[87] X. S. Si, C. H. Hu, and W. Wang, "An Adaptive and Non-Linear Drift-Based Wiener Process for Remaning Useful Life Estimation," in *Prognostics and System Health Management Conference (PHM-Shenzhen)*, pp. 1–5, 2011. (Cited on page 70.)

[88] W. R. Gilks, S. Richardson, and D. J. Speigelhalter, *Markov Chain Monte Carlo in Practice.* 1st edition, Chapman and Hall/CRC, 1996. (Cited on pages 72, 73, 74, and 75.)

[89] A. E. Gelfand and A. F. M. Smith, "Sampling Based Approaches to Calculating Marginal Densities," in *Journal of the American Statistical Association*, vol. 85, pp. 398–409, 1990. (Cited on page 72.)

[90] P. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations.* Springer, 1995. (Cited on page 73.)

[91] T. Hangelbroek and A. Ron, "Nonlinear Approximation Using Gaussian Kernels," in *Journal of Functional Analysis*, vol. 259, pp. 203–219, 2010. (Cited on page 75.)

[92] H. Akaike, "A New Look at the Statistical Model Identification," in *IEEE Transactions Automatic Control*, vol. 19, pp. 716–723, 1974. (Cited on page 77.)

[93] J. C. Lu and W. Q. Meeker, "Using Degradation Measures to Estimate a Time-to-Failure Distribution," in *Technometrics*, vol. 35, pp. 161–174, 1993. (Cited on page 77.)

[94] J. W. McPherson, "Reliability Challenges for 45nm and Beyond," in *43rd ACM/IEEE Design Automation Conference (DAC)*, pp. 176–181, 2006. (Cited on page 81.)

[95] S. Saini, *Low Power Interconnect Design.* Springer-Verlag New York, 2015. (Cited on pages 82 and 95.)

[96] M. R. Stan and W. P. Burleson, "Bus-Invert Coding for Low-Power I/O," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 3, pp. 49–58, 1995. (Cited on pages 82, 89, and 95.)

[97] C. Duan, Z. Chengyu, and S. P. Khatri, "Forbidden Transition Free Crosstalk Avoidance CODEC Design," in *45th ACM/IEEE Design Automation Conference (DAC)*, pp. 986–991, 2008. (Cited on pages 82 and 95.)

[98] C. Duan, V. H. C. Calle, and S. P. Khatri, "Efficient On-Chip Crosstalk Avoidance CODEC Design," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 17, pp. 551–560, 2009. (Cited on pages 82, 95, and 96.)

[99] D. Bertozzi, L. Benini, and G. de Micheli, "Error Control Schemes for On-Chip Communication Links: The Energy Reliability Tradeoff," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, pp. 818–831, 2005. (Cited on pages 82 and 96.)

[100] A. Sathish, M. M. Latha, and K. L. Kishore, "A Technique to Reduce Transition Energy for Data-Bus in DSM Technology," in *IJCSI International Journal of Computer Science.*, pp. 402–406, 2011. (Cited on pages 82 and 95.)

[101] J. Natesan and D. Radhakrishnan, "Shift Invert coding (SINV) for Low Power VLSI," in *Euromicro Symposium on Digital System Design (DSD)*, pp. 190–194, 2004. (Cited on pages 82 and 95.)

[102] J. Sathish and T. S. Rao, "Bus Regrouping Method to Optimize Power in DSM Technology," in *International Conference on Signal Processing, Communications and Networking (ICSCN)*, pp. 345–348, 2008. (Cited on pages 82 and 95.)

[103] C. L. Su, C. Y. Tsui, and A. M. Despain, "Saving Power in the Control Path of Embedded Processors," in *IEEE Design and Test of Computers*, vol. 11, pp. 24–30, 1994. (Cited on pages 82 and 95.)

[104] R. Murgai and M. Fujita, "On Reducing Transition Through Data Modifications," in *Design, Automation and Test in Europe (DATE)*, pp. 82–88, 1999. (Cited on pages 82 and 95.)

[105] R. B. Lin and M. Fujita, "Inter-Wire Coupling Reduction Analysis of Bus-Invert Coding," in *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 55, pp. 1911–1920, 2008. (Cited on pages 82 and 95.)

[106] L. Benini and et al., "Asymptotic Zero-Transition Activity Encoding for Address Buses in Low-Power Microprocessor-based Systems," in *7th Great Lakes Symposium on VLSI*, pp. 77–82, 1997. (Cited on pages 82 and 95.)

[107] L. Benini and et al., "System-Level Power Optimization of Special Purpose Applications: The Beach Solution," in *International Symposium on Low Power Electronics and Design (ISLPED)*, pp. 24–29, 1997. (Cited on pages 82 and 95.)

[108] M. R. Stan and W. P. Burleson, "Coding a Terminated Bus for Low Power," in *5th Great Lakes Symposium on VLSI*, pp. 70–73, 1995. (Cited on pages 82 and 95.)

[109] J. J. Benedetto and M. W. Frezier, *Wavelets: Mathematics and Applications.* (Cited on page 83.)

[110] R. Dennard, "Design of Ion-Implanted MOSFETs with Very Small Physical Dimensions," in *IEEE Journal of Solid State Circuits*, pp. 256–268, 1974. (Cited on page 95.)

[111] A. Ganguly and et al., "Addressing Signal Integrity in Networks on Chip Interconnects through Crosstalk-Aware Double Error Correction Coding," in *IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, pp. 317–324, 2007. (Cited on page 96.)

[112] R. Srinivasa and N. R. Shanbhag, "Coding for Reliable On-Chip Buses: Fundamental Limits and Practical Codes," in *18th International Conference on VLSI Design*, pp. 417–422, 2005. (Cited on page 96.)

[113] D. Rossi, C. Metra, A. K. Nieuwland, and A. Katoch, "New ECC for Crosstalk Impact Minimization," in *IEEE Design and Test of Computers*, vol. 22, pp. 340–348, 2005. (Cited on page 96.)

[114] K. N. Patel and I. L. Markov, "Error-Correction and Crosstalk Avoidance in DSM Busses," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 12, pp. 1076–1080, 2004. (Cited on page 96.)

[115] Chen, A., Hutchby, J., Zhirnov, V. and Bourianoff, G., *Emerging Nanoelectronic Devices.* Wiley, 2015. (Cited on page 100.)

[116] Morelos-Z., R. H., *The Art of Error Correcting Coding.* Wiley, 2006. (Cited on page 101.)

[117] G. Marsaglia, "Xorshift RNGs," *Journal of Statistical Software*, vol. 8, no. 14, 2003. (Cited on page 103.)

[118] "Code Tables: Bounds on the Parameters of Various Types of Codes.," 2015. `http://www.codetables.de/`. (Cited on page 107.)

[119] "Magma Computational Algebra System.," 2015. `http://magma.maths.usyd.edu.au/magma/`. (Cited on page 107.)

List of Publications

Publications related to the thesis

International Journals

1. N. Cucu Laurenciu, S.D. Cotofana, **Critical Transistors Nexus Based Circuit-Level Aging Assessment and Prediction**, Journal of Parallel and Distributed Computing 74(6), pp. 2512–2520 (2014).

2. N. Cucu Laurenciu, S.D. Cotofana, **A Nonlinear Degradation Path Dependent End-of-Life Estimation Framework from Noisy Observations**, Microelectronics Reliability 53(9–11), pp. 1213–1217 (2013).

3. N. Cucu Laurenciu, S.D. Cotofana, **Context Aware Slope Based Transistor-Level Aging Model**, Microelectronics Reliability 52(9–10), pp. 1791–1796 (2012).

International Conference Proceedings

1. N. Cucu Laurenciu, Y. Wang, S.D. Cotofana, **Haar-based Interconnect Coding for Medium/Long Range Energy Effective, Reliable Data Transport**, submitted to Design, Automation and Test in Europe (DATE), (2017).

2. N. Cucu Laurenciu, T. Gupta, V. Savin, S.D. Cotofana, **Error Correction Code Protected Data Processing Units**, Proceedings of the 12^{th} ACM/IEEE International Symposium on Nanoscale Architectures (NANOARCH), (2016).

3. N. Cucu Laurenciu, S.D. Cotofana, **Probability Density Function Based Reliability Evaluation of Large-Scale ICs**, Proceedings of the 10^{th} ACM/IEEE International Symposium on Nanoscale Architectures (NANOARCH), pp. 157–162 (2014).

4. N. Cucu Laurenciu, Y. Wang, S.D. Cotofana, **A Direct Measurement Scheme of Amalgamated Aging Effects with Novel On-Chip Sensor**, Proceedings of the 21^{st} IFIP/IEEE International Conference on Very Large Scale Integration (VLSI-SoC), pp. 246–251 (2013).

5. N. Cucu Laurenciu, S.D. Cotofana, **A Markovian, Variation-Aware Circuit-Level Aging Model**, Proceedings of the 8^{th} IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH), pp. 116–122 (2012).

Other publications

International Conference Proceedings

1. N. Cucu Laurenciu, S.D. Cotofana **Low Cost and Energy, Thermal Noise Driven, Probability Modulated Random Number Generator**, Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS), pp. 2724–2727 (2015).

2. A. Amaricai, N. Cucu Laurenciu, O. Boncalo, J. Chen, S. Nimara, V. Savin, S.D. Cotofana **Multi-Level Probabilistic Timing Error Reliability Analysis Using a Circuit Dependant Fault Map Generation**, Proceedings of Conference on Design of Circuits and Integrated Systems (DCIS), pp. 1–6 (2015).

3. A. Amaricai, V. Savin, O. Boncalo, N. Cucu Laurenciu, J. Chen, S.D. Cotofana **Timing Error Analysis of Flooded LDPC Decoders**, Proceedings of the IEEE International Conference on Microwaves, Communications, Antennas, and Electronic Systems (COMCAS), pp. 1–5 (2015).

Samenvatting

Het agressieve neerschalen van CMOS technologie tot in het tientallen-nanometer gebied, met het tegelijkertijd verbeteren van prestaties en yield, heeft de variabiliteit van device eigenschappen bepaald wat betreft hun nominale waardes, die tot grote spreiding kunnen leiden in delay, power, en robuustheid en daarnaast hun gevoeligheid voor ouderdom en door ruis veroorzaakte fouten tijdens normaal gebruik vergroot. Omdat de gate dielectric van transistoren de vermogensdichtheid en elektrische veldsterkte doen toenemen, zijn ook de factoren die faalmechanismen van geïntegreerde schakelingen (IC - Integrated Circuits) versnellen ernstiger dan ooit geworden, wat kan resulteren in hogere uitvalpercentages tijdens de levenscyclus van ICs en eerdere start van het verouderingsproces. Het gevolg hiervan is dat het behalen van de betrouwbaarheidsdoelen binnen haalbare kosten in dit landschap een significante uitdaging wordt, die op een unitaire manier moet worden benaderd van ontwerp tot run-time. Om dit te bewerkstelligen stellen wij een holistisch, betrouwbaarheids-bewust ontwerp- en levencyclusbeheerframework voor, dat bestaat uit een design-time en run-time gedeelte. Het resultaat van het design-time gedeelte van het framework is een adaptieve fysieke structuur met verbeterde betrouwbaarheid. Het run-time gedeelte is in staat om het slijtageprofiel van deze structuren te observeren en dynamisch te beheren zodat aan gebruikergedefinieerde servicekwaliteitseisen (Quality-of-Service) zal worden voldaan. Ook kan het een volledig logboek bijhouden van de betrouwbaarheid om als hulpinformatie gebruikt te kunnen worden bij het ontwerp van de volgende generatie ICs. Specifiek introduceren wij ten eerste ouderdomsmodellen die tijdens design-time op transistor- en circuit-niveau kunnen worden gebruikt, en die een basis leggen voor 4-dimensionale Design-Space-Exploration (DSE) die een betrouwbaarheid-geoptimaliseerd circuit realisatie kan identificeren dat aan bepaalde eisen voor circuit oppervlakte, power, en delay voldoet. Vervolgens, om het creëren van een low-cost maar toch nauwkeurige infrastructuur voor het observeren van de fysieke structuren mogelijk te maken, stellen we een methodologie voor om het aantal ouderdomssensoren te minimaliseren and om hun optimale locatie te identificeren, en we introduceren een sensorontwerp dat in staat is om op circuit-niveau direct gecombineerde effecten van gelijktijdige degradatiemechanismen te detecteren. Verder, om de informatie die vanuit sensors verzameld is zinvol te

maken voor het run-time beheerframework, introduceren we een model dat de globale veroudering van het circuit kan schatten en haar End-of-Life kan voorspellen. Dit gebeurt op basis van onnauwkeurige sensormetingen en er wordt rekening gehouden met de non-lineariteit van het degradatieproces. Ten slotte, om bij de DSE meer opties te bieden voor de verbetering van de betrouwbaarheid, richten we ons op het realiseren van betrouwbare datatransport en -verwerking met onbetrouwbare componenten. Hiervoor introduceren we: 1) een codec voor betrouwbare energie-efficiënte middellange/langeafstand datatransport, en 2) een methodologie om eenheden te ontwerpen die door Error Correction Codes beschermde data kunnen verwerken met een kleinere error rate dan die van de fabricatietechnologie.

Propositions

accompanying the PhD dissertation

Reliability Aware Computing Platforms

Design and Lifetime Management

---

1. A holistic framework able to deal with IC reliability issues in a coherent manner from design to run time is crucial for enabling dependable computing. [This Thesis]
2. Understanding device aging is the only avenue to a prolonged IC lifetime. [This Thesis]
3. Transistors are relevant for reflecting circuit aging, not as individuals but as members of a crowd. [This Thesis]
4. Reliable data processing units can be obtained by intimately intertwining error correcting codes with processing unit functionality. [This Thesis]
5. Any non-oligarchic political system eventually devolves into an oligarchy, the difference resides in the RC timing constants.
6. Character and intelligence should come as a package, however, intelligence is frequent, while character is seldom.
   [Mihai Eminescu - Romanian poet (1850-1889)]
7. "Omnia in opinione sita."
   [Marcus Aurelius - Roman emperor and stoic philosopher (121-180)]
8. You may have to fight a battle more than once to win it.
   [Margaret Thatcher - British politician (1925-2013)]
9. The longest road is the one towards yourself.
   [Nicolae Iorga - Romanian historic (1871-1940)]
10. Sunny day, happy day, happy life.

These propositions are regarded as opposable and defendable, and have been approved as such by the promotor, Prof. dr. K.L.M. Bertels.

Curriculum Vitae

Nicoleta CUCU LAURENCIU was born on 3^{rd} of March 1986 in Brasov, Romania. She received the B.Sc. degree in Electrical Engineering and Computer Science, Electronics Specialization, from "Transilvania" University of Brasov, Romania, in 2009, and the M.Sc. degree in Computer Engineering from Delft University of Technology, Faculty of Electrical Engineering, Mathematics and Computer Science, The Netherlands, in 2010. Subsequently, she started in 2011 the Ph.D. studies, at the same faculty and departament, at Delft University of Technology, under the supervision of Dr. S. D. Cotofana. The focus of her Ph.D. studies is on dependable computing and architectures based on unreliable components.