Digital circuit performance estimation under PVT and aging effects

Mauricio Altieri Scarpato

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTE UNIVERSITE GRENOBLE ALPES

Spécialité : **Nano Electronique et Nano Technologies**

Arrêté ministériel : 25 mai 2016

Présentée par

Mauricio ALTIERI SCARPATO

Thèse dirigée par **Édith BEIGNÉ, HDR, CEA-LETI**, et codirigée par **Suzanne LESECQ, HDR, CEA-LETI**

préparée au sein du **Laboratoire d'Electronique et des Technologies de l'Information, CEA Grenoble**

dans **l'École Doctorale Electronique, Electrotechnique, Automatique et Traitement du Signal (EEATS)**

Estimation de la performance des circuits numériques sous variations PVT et vieillissement

Thèse soutenue publiquement le **12 décembre 2017**, devant le jury composé de :

Mme. Lorena ANGHEL
Grenoble INP, Présidente
M. Daniel MENARD
INSA Rennes, Rapporteur
M. Abdoulaye GAMATIE
LIRMM, Rapporteur
M. Olivier HÉRON
CEA-LIST, Examinateur
Mme. Édith BEIGNÉ
CEA-LETI, Directeur de thèse, Invitée
Mme. Suzanne LESECQ
CEA-LETI, Co-directeur de thèse, Invitée

Acknowledgments

It is my pleasure to acknowledge the roles of several individuals who were fundamental for completion of my PhD thesis.

Firstly, I would like to express my gratitude to Abdoulaye Gamatié and Daniel Ménard for reviewing my manuscript and giving me important feedbacks. My sincere thanks also to Lorena Anghel, whom I had the pleasure of having as teacher at ENSIMAG, for accepting the invitation to chair my thesis defense.

I am highly indebted and thoroughly grateful to my supervisors Edith Beigné, Olivier Héron and Suzanne Lesecq who encouraged and directed me through these three years. This thesis would not come to existence without their guidance and large background in different specialties. More than anything, it was their strong support and confidence on me that kept me up during the moments of uncertainty of the thesis.

I am grateful to all my colleagues from both LIALP and LISAN laboratories. Having such amazing work environment was essential for succeeding in my PhD. I must confess that the almost daily croissants/cakes in the cafeteria were an extra motivation to get to the lab every morning. I especially mention Adja Sylla and Thierno Barry who shared with me this hard but rewarding experience of doing a PhD. I am thankful to Vincent Olive, head of LIALP, as well as to Fabien Clermidy and Jérôme Martin, former and current heads of LISAN, for providing me with all meanings I needed to make the most of my time as PhD student at CEA.

I am also deeply thankful to Ivan Miro-Panades who shared with me his vast knowledge on digital monitors and circuit reliability. Thanks to David Coriat for providing me the netlists from the FRISBEE circuit. Thanks to Anca Molnos for helping me with learning methods. Thanks to Diego Puschini for having mentored me in various subjects since my internship and for bringing *mate* to the lab. Thanks to Marc Belleville and Pascal Vivet for their advices. Thanks to Chiara Sandionigi as well as to all LCE co-workers who I had the pleasure to meet in Saclay. I had a short but enriching stay there. Finally, thanks to everyone who helped me directly or indirectly during my PhD, it would not be

possible without you.

My deepest thanks for all my longtime friends from Brazil and the new friends I met in Grenoble. They all make this long journey considerable easier by sharing unforgettable moments with me. A special thanks to my roommates Arthur and Joo Paulo for making our *coloc* the best one in Grenoble. I want to extend my thanks to Fernanda Kastensmidt and José Rodrigo Azambuja for introducing myself to the research environment at UFRGS.

Lastly but most importantly, I want to thank my family for all support they gave me despite the distance. My family have always been my base of support, even more so after I came to France. They always encouraged me to pursue my dreams even though it meant that I would be far away from them. Most of all, I dedicated this thesis and all my achievements to my mother who made my happiness the priority of her life since I was born. *Obrigado mãe, te amo muito!*

Contents

List of Acronyms

ABB Adaptive Body Bias 20
ADC Analog-to-Digital Converter 29
AFS Adaptive Frequency Scaling v
AVFS Adaptive Voltage and Frequency Scaling 17
AVS Adaptive Voltage Scaling 18
BTI Bias Temperature Instability 1
CPR Critical Path Replica 23
DVFS Dynamic Voltage and Frequency Scaling 17
EDA Electronic Design Automation 16
FIT Failure in Time xi
FLL Frequency-Locked Loop 19
HCI Hot Carrier Injection 1
LUT Look-Up Table 90
PLL Phase-Locked Loop 19
PVT Process, Voltage and Temperature 1
PWM Pulse Width Modulator 18
RISC Reduced Instruction Set Computer 80
SSTA Statistical Static Timing Analysis 16
STA Static Timing Analysis 16
UTBB FD-SOI Ultra-Thin Body and Box Fully-Depleted Silicon-On-Insulator 21
VCO Voltage-controlled oscillator 19
Vbb Body bias voltage 20
Vdd Supply voltage x
Vth Threshold voltage 7

Introduction

"With engineering, I view this year's failure as next year's opportunity to try it again. Failures are not something to be avoided.
You want to have them happen as quickly as you can so you can make progress rapidly."

Gordon Earle Moore,
Intel co-founder and author of Moore's law.

Context

In 1965, Gordon Moore forecast that the number of transistors in an integrated circuit would double every two years [1], which has since become widely known as Moore's Law. The continuous increase in transistor density has lead to huge increases both in performance and in the number of functionalities per circuit. In turn, this has boosted the development of mobile applications like laptops, tablets and mobile phones. Besides calling and texting, current mobile phones incorporate additional features like browsing, playing games, taking pictures, etc... Moreover, a phone nowadays has the same processing performance than a high-end desktop computer had ten years ago.

Due to physical limitations (weight and size), the battery capacity of such devices cannot be expanded at the same rate as their performance. For instance, the Samsung Galaxy S5 phone, launched in 2014, had a battery capacity of 2800 mAh. Three years later, the Samsung Galaxy S8 phone was launched with a battery of 3000 mAh. This corresponds to an increase of only 7% in battery capacity while the processing performance has increased about 300% [2]. More than the performance, it is therefore the energy efficiency that must be improved, also known as performance per watt. Otherwise, the battery would not be able to afford such processing power.

Improvements on energy efficiency are mandatory not only for mobile applications. The concept of Internet of Things (IoT) has become popular in the last years. It corresponds to devices that interoperate through network connectivity. This leads to new application domains, for instance smart house, healthcare, defense, industry automation and transport network, where the interconnection of

theses devices provide new services to the end-users. Intel has estimated that 2 billion devices were connected by 2006 while this number raised to 15 billions by 2015. At that pace, they expect 200 billions connected objects by 2020 [3]. With such an escalation in the number of connected devices, it is no wonder that their energy consumption has become a major issue. Indeed, a report issued in 2015 by the Semiconductor Industry Association (SIA) stated that the computer devices will require more energy than the world can generate by 2040 [4].

Since both dynamic and leakage power depend on the Supply voltage (Vdd), the main approach to reduce energy consumption so far consisted in scaling down Vdd with transistor size. However, the reduced transistor dimensions has exacerbated the impact of variability on CMOS circuits. The sources of variability can be either static (due to manufacturing imperfections - P variability) or dynamic (due to Vdd and temperature fluctuations - VT variability). Therefore, deep submicron technologies require the use of large voltage guard-bands to ensure a "correct" operation of the circuit under these different sources of variability, the objective being to have a functional circuit even in the presence of PVT variability. This has lead to a slowdown of Vdd scaling in recent technology nodes, as shown in Figure 1 [5].

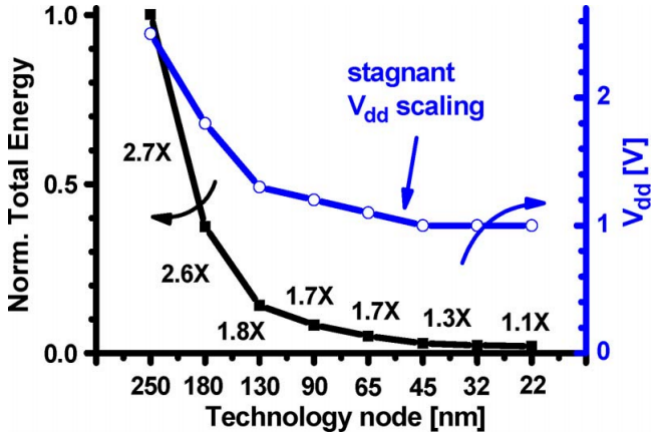


Figure 1: Energy and Vdd scaling vs. technology node [5].

Considerable energy savings can be achieved by decreasing these safety margins. This is done through the use of an adaptive management scheme. Such techniques use embedded sensors that track the fluctuation of the circuit timing induced by static and dynamic variations. Then, a variable voltage actuator and/or clock frequency actuator (and possibly a body bias actuator) changes on the fly the operating conditions of the circuit to reduce its energy consumption while avoiding timing faults. Vdd is kept at its minimum functional value or, respectively, the clock frequency is kept at its maximum value. The main limitation of current adaptive techniques is that they do not distinguish the source of

the variation while they detect the variation of the circuit timing. In advanced technology nodes, aging has become an important source of variability, in particular BTI and HCI effects. Unlike other dynamic variations, the aging-induced shift of the circuit performance is permanent. It can thus lead the circuit to an irreversible unreliable condition.

Aging might not be a real problem for customer electronics, because their life span is smaller than 10 years and they usually do not operate at harsh environments. Nevertheless, it is a major issue for safety-critical real-time systems, such as aerospace and automotive ones. These systems demand high performance with very low failure rate. For example, avionics systems require over 25 years of operation with a maximum failure rate of 100 Failure in Time (FIT) [6], where 1 FIT corresponds to 1 failure per billion hours. Moreover, they are used in extreme conditions, where the temperature can range from -50$°C$ to 150$°C$. The mission profiles of such critical systems are highly diversified making it impossible to accurately estimate the aging degradation during the design phase. The worst-case scenario approach is thus taken into account to fulfill the performance and safety requirements. However, this leads to an excessive energy consumption due to the pessimistic voltage guard-bands.

Motivations and Objective

Aging of human beings is determined by the way we live, i.e. how healthy we eat, how much sport we do, etc... This is also valid for an electronic circuit whose aging does not depend only on its age. Two identical circuits with the same age do not necessarily present the same aging degradation. The degradation depends on how the circuit was used during its lifetime, i.e. the historic of its operating conditions, namely, supply voltage, temperature and workload. This is why the estimation of circuit aging is not simple. Moreover, the impact of aging on the circuit performance is governed by the operating condition once the degradation of the transistor characteristics causes a shift in the propagation delay that depends on the PVT conditions. For instance, while a transistor degrades faster at a higher supply voltage, an aged transistor has a higher relative influence on the circuit performance when it operates at a lower supply voltage.

In the last years, many works have focused on modelling aging effects in transistors, particularly BTI and HCI. Such works managed to provide accurate models for the aging-induced shift of transistor characteristics, especially the threshold voltage. These models are based on the physical mechanisms of aging and they are validated on silicon data for several combinations of voltage and temperature values. Such models can be integrated in electronic circuit simulators, e.g. SPICE, which allows the assessment of the circuit degradation for any operating condition. However, as stated above, the mission profile of a circuit is

barely known at the design phase. The use of existing aging models is thus not enough for estimating the circuit degradation, whatever their accuracy is.

Having an on-line estimation of the circuit health would allow to calculate on the fly the required safety guard-bands. Besides, it could be used by reliability strategies to improve the circuit lifetime. However, an integrated circuit nowadays is composed of millions or billions of transistors. Even a single critical path comprises more than a hundred transistors. Applying transistor-level models to on-line estimate the circuit degradation is computationally impractical. AS a consequence, simplified models must be adopted so that the computation overhead is not larger than the benefits provided by its use.

The objective of this PhD thesis is to propose a methodology to abstract the complexity of existing aging models. The idea is basically to generate circuit-level aging models for synchronous digital circuits. Such models are simple equations that provides the propagation delay of a critical path considering PVT variations together wigh aging. The historical values of PVT are taken into account to estimate the aging degradation. Instead of estimating the V_{th} shift for each transistor, the proposed models estimate an overall parameter shift for the whole path. Being architecture- and technology-independent, this methodology can be applied to any digital circuit.

Contributions

The main contributions of this work are:

- **Propose a methodology to generate a circuit-level aging model from device-level models**

  A methodology is proposed to model the propagation delay of a critical path from device-level models. The resulting model gives the critical path delay based on the PVT conditions. The parameters of the model are obtained through non-linear regression from data obtained with SPICE simulations.

  From the propagation delay model, a second model for both BTI and HCI effects is proposed. The model reflects the aging-induced propagation delay shift for any PVT condition. In other words, it is integrated in the delay model as a parameter shift. This latter model takes into account all factors that influence the aging degradation, namely, supply voltage, temperature, workload and circuit topology. Dynamic variations are also taken into account to allow on-line estimation.

  This contribution has been published in "Towards on-line estimation of aging in digital circuits through circuit-level models", IRPS'17 [7]. A patent application, "Method and device for estimating circuit aging" [8], has also been filed.

- **Develop an aging-aware solution to estimate voltage and temperature variations**

  Many sensor solutions exist in the literature to track voltage and temperature changes. However, none had directly addressed the impact of aging effects on its operation so far, although some works assume that their architecture is robust to aging. Moreover, an aging-aware VT sensor must be adopted to allow the use of the proposed circuit-level models to on-line estimate the circuit health.

  Therefore, we first analyzes the impact of both BTI and HCI effects on the voltage and temperature estimates provided by a small area digital sensor. Then, we proposed a recalibration method that increases the sensor robustness against aging effects.

  This work has been presented in the publication "Evaluation and mitigation of aging effects on a digital on-chip voltage and temperature sensor", PATMOS'15 [9].

- **Demonstrate the application of the proposed models in different contexts**

  The circuit-level models proposed in this thesis can be used either on-line to estimate the circuit health or off-line to simulate the operation of the circuit under aging effects. We shortly discuss four possible on-line applications of the proposed models. The proposed applications consist in an adaptive control, a dynamic Mean Time To Failure (MTTF) computation, an estimation of the maximum operating conditions for a given MTTF and a reliability measure for task mapping in multi-core circuits. Then, the models are used in two off-line modelling contexts:

  - the first one is a framework to simulate a multi-core circuit. This framework allows the implementation of strategies of task mapping and DVFS. Different strategies are implemented and evaluated here with respect to aging, performance and energy consumption;

  - the second is an adaptive system implemented as a Simulink model. This circuit constantly changes its clock frequency based on embedded sensors that are placed at the critical paths and warns the pre-occurrence of a timing fault. From the simulations, a technique is proposed to estimate the aging-induced performance shift of such systems by only tracking temperature variations.

  The latter contribution is presented in the publication "Tracking BTI and HCI effects at circuit-level in adaptive systems.", NEWCAS'16 [10].

Report Organization

Apart from this introduction, the report is composed of four chapters and a conclusion. Chapter 1 introduces the sources of variability in advanced CMOS nodes and their impact on digital circuits. Firstly, it presents the process variations due to manufacturing and atomistic limitations. These variations lead to different values of performance and power than those estimated at the design phase. Next, the dynamic variations are addressed. Local variations of voltage and temperature depend on the circuit operation and they have a strong impact on the switching speed of the transistors. Finally, aging-induced variations are presented. Aging has become a major issue in recent technology nodes, in particular BTI and HCI effects. These phenomena cause parametric shifts in transistors which may result in a faulty operation of the circuit. This chapter exposes then the importance and need for correct modelling of variability as well as adaptive techniques.

Chapter 2 firstly explains how the variability is estimated during the design phase of a circuit and why the traditional approach of safety guard-banding is not anymore suitable. Then, it introduces the concept of adaptive circuit which is basically a circuit that implements an adaptation strategy based on a monitoring system. The adaptation strategy consists in changing the operating conditions of the circuit to cope with variability. This adaptation can be done through the supply voltage and/or the clock frequency. Next, this chapter highlights some existing monitors to track variability. The presented solutions range from timing fault detection to direct measure of the variation. It is shown that current solutions are not able to provide an accurate information about the aging-induced performance degradation. Finally, this chapter addresses the impact of aging on the voltage and temperature estimation with a digital sensor. A recalibration method is proposed to mitigate the aging effects on the voltage and temperature estimates.

Chapter 3 presents the main contribution of this work. It consists in a methodology to construct circuit-level aging models from device-level models. Based on SPICE simulations, a first model is generated for the propagation delay of a critical path. The resulting model is an equation that depends on the Process-Voltage-Temperature variability. Next, both BTI and HCI effects are modelled as a shift in the parameters of the previous equation. The proposed model takes into account all factors that impact aging, namely, circuit topology, supply voltage, temperature and workload. The proposed methodology is validated on two different architectures implemented in $28nm$ FD-SOI technology.

Chapter 4 demonstrates the use of the proposed models in different contexts. First, four on-line applications are shortly discussed. The applications range from the integration in an adaptive system to a dynamic Mean Time to Failure computation. Then, the models are used for a multi-core simulation framework. This framework allows the evaluation of different task mapping strategies with

respect to reliability, power and performance. In the end, the operation of a complete AFS system is modelled in Simulink with the help of the proposed models. As a result, a simplified method is developed to estimate the circuit degradation by tracking both the clock frequency and the temperature variations over time.

Finally, Chapter 5 summarizes the main contributions and proposes future work directions.

Chapter 1

Sources of variability in advanced technology nodes

The continuous shrinking of transistor size leads to great advances in circuit performance besides reducing energy consumption and transistor cost. However, this aggressive scaling also makes the CMOS circuits more susceptible to variability. There exist 3 main sources of variability, namely, Process, Voltage and Temperature (PVT) variations. Process variations are due to the mismatch of the manufacturing process. Voltage variations are mostly due to the parasitic impedance while temperature variations are caused by the power dissipated by the circuit. PVT variations change the transistor switching speed and leakage current. These variations may uniformly impact the whole circuit (i.e. global variations) or impact each part of the circuit in a different way (i.e. local variations).

Furthermore, aging effects emerged as a new source of variability in recent technology nodes. Both Bias Temperature Instability (BTI) and Hot Carrier Injection (HCI) effects degrade the circuit performance by increasing the transistor threshold voltage over time. All this increase of variability is transforming the circuit design in a probabilistic problem instead of a deterministic one. Large safety guard-bands are thus necessary to guarantee a correct operation under the worst-case conditions. In this chapter, we summarize the origin of each source of variability and its impact on digital CMOS circuits.

This chapter is organized as follows. Section 1.1 describes the sources of process variations at both global and local hierarchical level as well as their consequences for digital circuits. Next, Section 1.2 and Section 1.3 present the dynamic environmental variations, namely, supply voltage and temperature, respectively. Finally, Section 1.4 talks about aging effects, in particular BTI and HCI.

1.1 Process variations

Process-induced variations have always been a key concern in integrated circuit design. Process variations imply in the mismatch of transistors characteristics from nominal values. This leads to values of propagation delay and current leakage different from those estimated during the design phase resulting in yield loss. Historically, process variations were mainly due to manufacturing process imperfections. But, as the transistor size approaches the atomic scale, variations due to the intrinsic atomistic nature become more important.

Two types of process variability exist, namely systematic variations and random variations. Systematic variations are deterministic and repeatable deviations that depend on the spatial position of the transistor on the wafer and on its surrounding layout. Systematic effects are mainly due to photolithography limitations. Deep ultra-violet laser with $193nm$ wavelength is still the main light source used in semiconductor lithography even though technology nodes as small as $10nm$ are already in manufacturing. Variations due to sub-wavelength lithography occur from the correlation between adjacent structures, such as lithography proximity effects (LPE) and well proximity effects (WPE). Since these effects follow a clear pattern and are layout-dependent, they can be modeled, predicted and even corrected through resolution enhancement technologies (RETs) such as optical proximity correction (OPC) and phase shift mask (PSM) [11].

On the other hand, random variations are a stochastic phenomenon without clear patterns. Random variations are mainly due to atomistic limitations and they have become the dominant source of process variation as the transistor size scaled below $90nm$. Some types of random variations are random dopant fluctuation (RDF), line edge roughness (LER) and gate thickness fluctuation (OTF) [12]. As it is not possible to predict these variations, they require a different statistical treatment. Typically, random variations are characterized and modeled as a Gaussian distribution. Circuits are then designed to reach a given yield considering this probability function. Note that higher the yield desired, higher the safety margins needed.

1.1.1 Global process variations

Besides systematic and random natures, process variations are also classified between global and local. Global variability corresponds to deviations of physical parameters from nominal values with the same change for every transistor on a die. These parameters include channel length (L), channel width (W), layer thickness, resistivity, doping density and body effect [12]. As a consequence, identically designed circuits may have different performances and power characteristics. As shown in Figure 1.1 [13], the performance difference between dies within a wafer can be up to 30% while the leakage current can vary up to a factor

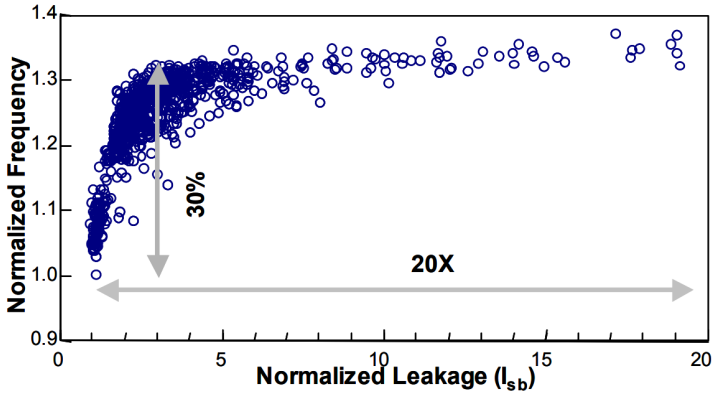


Figure 1.1: Frequency and leakage current distribution between dies within a wafer fabricated in 180nm CMOS technology[13].

of 20.

Global variations can be separated in lot-to-lot (L2L), wafer-to-wafer (W2W) and die-to-die (D2D). A lot usually contains 25 wafers, while a wafer may contain hundreds or thousands of dies depending on the die size. Global variability is mostly systematic and, thus, spatially and temporally correlated. This means that dies that are closer to each other in a wafer are more prone to have undergone the same variability. The same is valid for lots manufactured over a short period of time.

In order to take global variations into account during design, foundries perform device characterization through I-V measurements, providing designers with models for best- and worst-case of transistor parameters in addition to nominal values. *Fast/fast (FF)* and *slow/slow (SS)* corners correspond to the best and worst values for PMOS and NMOS characteristics, respectively. They are represented by three standard deviations from the nominal values. The FS and SF corners are called as "skewed" corners. They are considered a concern for analog circuits, but are of a minor importance in digital designs [14].

A circuit is usually designed taking into account the worst-case scenario, i.e. with slow NMOS and slow PMOS transistors. However, this approach leads to a considerable energy loss since either a lower clock frequency than the maximum one or a higher supply voltage than the minimum one has to be used. Some companies adopt *speed-binning* by testing and separating the fabricated chips according to their maximum operating frequency. Yet, individually testing each circuit after fabrication is time- and cost-consuming that it is rarely worth doing.

1.1.2 Local process variations

Local variability, also called within-die (WID) variability, makes identically sized transistors placed in the same die have different electrical characteristics such as

threshold voltage. Despite having a systematic part, local variations are dominated by random components and are thus spatially uncorrelated [15]. Random dopant fluctuation (RDF) is the main source of local variation and is caused by the mismatch in the amount of dopants in the channel. Note that the number of dopant atoms in the channel reduces as the technology scales down, as shown in Figure 1.2 [16]. As a consequence, the reduced number of dopants considerably increases the impact of RDF on the threshold voltage variation since each mismatched atom has a larger importance.

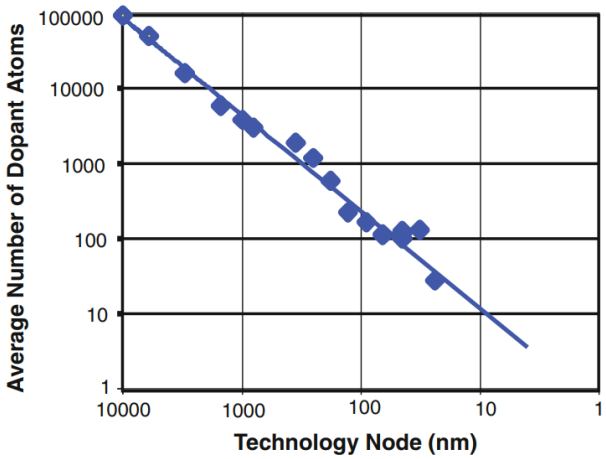


Figure 1.2: The number of dopant atoms per transistor is reduced with technology node [16].

Other sources of local random variations are line edge roughness (LER) [17] and oxide thickness fluctuation (OTF) [18]. LER is the deviation of the gate shape from an ideal smooth edge. It was not a concern in past technologies since the transistor dimensions were much more larger than the roughness. However, LER has not scaled down with technology becoming thus an important source of variation in technology nodes below $40nm$. OTF is induced by the atom-level interface roughness between silicon and gate dielectric. Similarly to LER, it has also become a major issue in recent technology nodes as the oxide thickness is reduced to only a few silicon atomic layers.

Local random threshold voltage variation is usually modelled as a stochastic process with a standard deviation $\sigma_{V_{th}}$. $\sigma_{V_{th}}$ is reported to be inversely proportional to the gate area, as follows [19]:

$$\sigma_{V_{th}} \propto \frac{1}{\sqrt{WL}} \tag{1.1}$$

where W and L are the width and the length of the transistor gate, respectively. Figure 1.3 shows the $\sigma_{V_{th}}$ evolution with respect to technology node [20] for the three main sources of local random variation. As the threshold voltage has an

important role on the switching speed of the transistors, an increase in $\sigma_{V_{th}}$ will result in an increase in the path delay uncertainty.

The traditional approach to estimate local variations during circuit design is to perform heavy Monte Carlo simulations including global variations as well. This provides a statistical analysis of both circuit performance and yield under all sources of process variation. But, as it can be seen through Figure 1.3 and equation (1.1), it becomes harder to handle local variations as the transistor size scales down. Increasing voltage guard bands are thus mandatory to reach an acceptable yield which, in turn, leads to poor energy efficiency. This issue will be better covered in the next chapter.

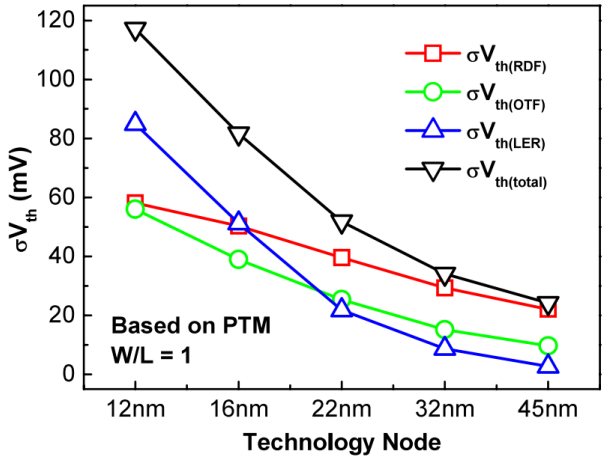


Figure 1.3: Local threshold voltage variation $\sigma_{V_{th}}$ induced by random dopant fluctuation (RDF), line edge roughness (LER) and oxide thickness fluctuation (OTF) versus technology node [20].

1.2 Supply voltage variations

In addition to static process variations, the circuit is also affected by dynamic environmental variations, namely, supply voltage and temperature variations. Supply voltage variations are mainly caused by voltage drop and current derivative di/dt noise [14]. Voltage drop, also called IR drop, occurs when the current flows over the parasitic resistance of the power grid while di/dt noise is caused by the parasitic inductance. As packaging and platform technologies do not progress at the same rate than CMOS process, the circuit impedance does not scale as fast as the supply voltage. This leads to an influence of voltage variations relatively more important as the technology scales down [13].

The duration of a voltage fluctuation ranges from nanoseconds to microseconds. As shown in Figure 1.4, voltage variations can be classified in 3 categories.

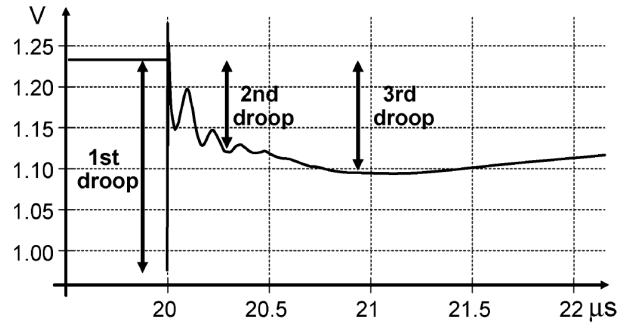


Figure 1.4: Three different timing categories of voltage drop [21].

The *first droop* lasts only a few nanoseconds but it is usually the deepest voltage drop. It is due to both the package inductance and the on-die capacitance and it can lead to a timing fault if a critical path is activated during the drop. The *second droop* depends on the package decoupling. It has a duration of some hundred nanoseconds. Finally, the *third droop* persists for a few microseconds. However it can be reduced through the use of bulk capacitors. After a voltage drop, the supply voltage always return to its nominal value until a new drop occurs.

A voltage drop can lead to variations up to 10% of the nominal voltage. The consequence of a voltage variation over the circuit delay depends on the supply voltage itself, as shown in Figure 1.5 for a 5 mV voltage drop. Lower the supply voltage, larger the delay variation. This comes from the fact that the transistor switching time is determined by its drain current which, in turn, depends on the

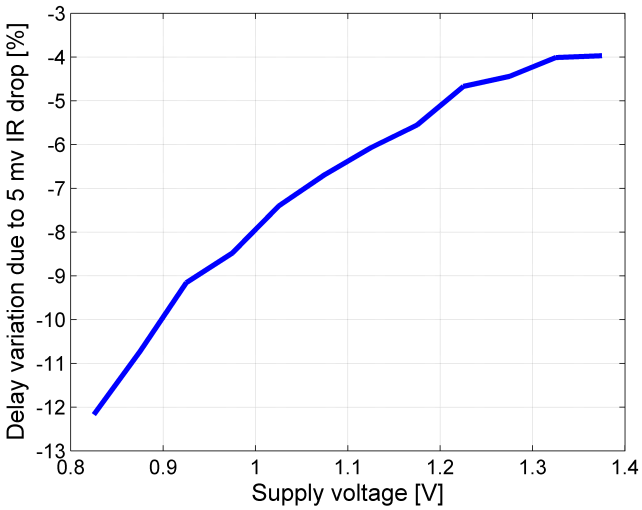


Figure 1.5: Path delay variation in percentage to a 5 mV voltage drop depending on the supply voltage.

difference between the V and the threshold voltage Threshold voltage (Vth). In [22], the authors demonstrated that the propagation delay of a logic gate, e.g. an inverter, is proportional to the supply voltage as follows:

$$t_{switch} \propto \frac{V}{(Vth - V)^{\alpha}} \tag{1.2}$$

where α is a technology-dependent factor. As it can be seen, the importance of a voltage variation ΔV increases as V gets close to Vth. Finally, voltage variations are not uniform around the die, they are spatially correlated. A voltage drop propagates through the power grid affecting each circuit logic block in a different way depending on its spatial position [23]. Handling voltage variations in recent technology nodes requires the use of local sensors or else large voltage guard-bands.

1.3 Temperature variations

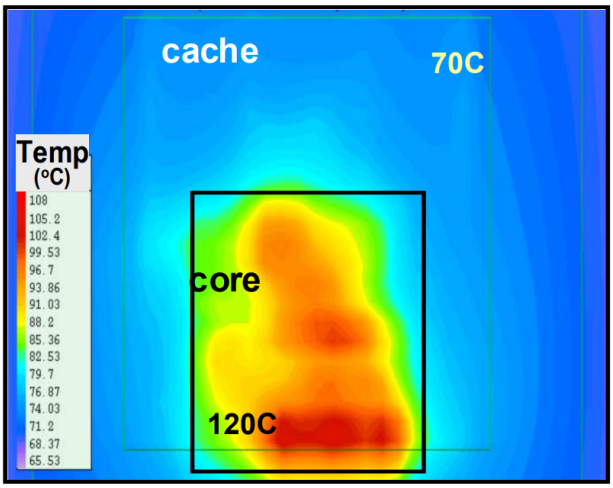


Figure 1.6: Temperature variation within a chip due to hot-spots [13].

Another source of dynamic variation in CMOS circuits is the temperature. Besides fluctuations of the ambient temperature, a circuit is also susceptible to local temperature variations induced by the power dissipated by its transistors. Since an electronic circuit does not perform any chemical or mechanical work, almost all the electrical energy consumed is transformed in thermal energy. The temperature inside a circuit can vary more than $30°C$ depending on the spatial position, as shown in Figure 1.6. This internal variation is due to *hotspots*, i.e., regions of the circuit that have a high activity and, consequently, dissipate more power. Temperature variations are slower than supply voltage ones since they have time constants ranging from miliseconds to seconds [24].

While an elevated temperature reduces the interconnect performance, the effect of temperature variations on the propagation delay of logic gates depends on the supply voltage. An increase in temperature will increase the transistor switching speed at a low value of V while the inverse is observed at a high value of V. This phenomenon is called the inverse temperature dependence (ITD). It is due to the fact that a higher temperature will reduce two transistor parameters, namely, threshold voltage and carrier mobility. Yet, these two parameters impact the transistor switching speed in opposite ways. A reduced threshold voltage leads to faster transitions while a smaller carrier mobility leads to slower transitions. As can be seen in Figure 1.7, the path propagation delay increases with the temperature for values of V above 0.95V while it decreases for values of V below 0.95V.

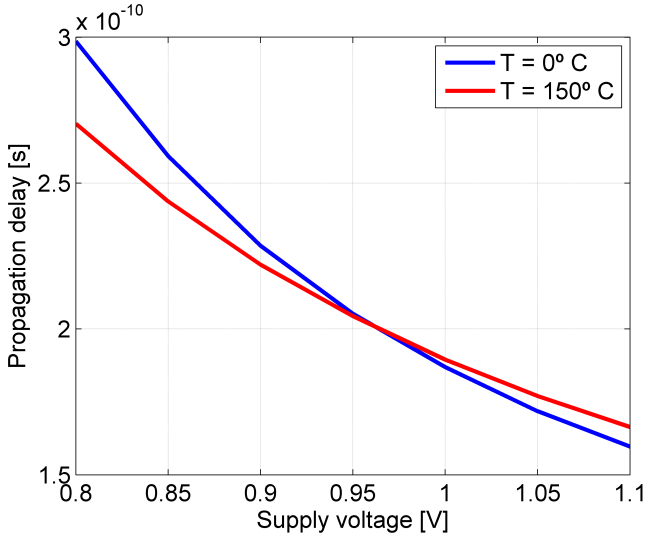


Figure 1.7: Demonstration of the inverse temperature dependence (ITD). A higher temperature increases the path propagation delay for values of V higher than 0.95V while the inverse is observed for V lower than 0.95V.

The circuit performance is not the only issue related to the temperature. By reducing the threshold voltage, the increase in temperature also leads to more leakage power dissipated by the transistors. Moreover, it has a direct influence on the aging-induced circuit degradation as explained in the next section. Finally, as the technology scales down, the transistor density in a circuit considerably increases. This, in turn, causes the elevation of the power density and, therefore, of the self-heating effect. Due to thermal design power constraints, parts of the circuit must be thus powered-off, the so-called dark silicon. Recent studies claimed that the amount of dark silicon may reach up to 80% of the circuit with $8nm$ CMOS technology [25].

1.4 Aging effects

Actually, aging effects on CMOS transistors is not a totally new topic in microelectronics. Negative Bias Temperature Instability (NBTI) has been reported for the first time in 1966 [26]. However, it was not until recently that aging has become a real issue in CMOS circuits, when gate oxide thickness scaled to values lower than $1.5nm$ [27, 28]. A smaller gate oxide thickness leads to a higher oxide electric field which is the dominant factor in the main aging phenomena such as NBTI, HCI and TDDB.

1.4.1 Negative/Positive Bias Temperature Instability (N/PBTI) and Hot Carrier Injection (HCI)

Negative Bias Temperature Instability (NBTI) consists in the degradation of PMOS transistors due to charges that get trapped inside the dielectric. These charges alter the transistor parameters over time, particularly the threshold voltage (Vth). Vth is increased due to the trapped charges reducing the transistor switching time. NBTI is claimed to be the most prominent aging phenomenon in digital CMOS circuits [29]. Positive Bias Temperature Instability (PBTI) is the equivalent degradation in NMOS transistors. However, PBTI effect is considerably smaller than NBTI and it only became an important degradation mechanism after the emergence of High-κ/Metal Gate (HKMG) transistors [30].

BTI degradation is reported to be composed of two mechanisms [31], namely, a recoverable part and a permanent one. The recoverable degradation is due to charge-trapping in preexisting traps inside the dielectric. The channel charges get trapped in the dielectric when a voltage is applied on the gate and they are released to the channel when the voltage is removed. The permanent degradation is called interface state generation because it is caused by the break of Si-H bonds located at the silicon-oxide interface. Both mechanisms have an exponential dependence on the oxide electric field, determined by the supply voltage, and on the temperature [31]. An increase in temperature reduces both capture and release time of charges inside the oxide [32]. In other words, transistors get more degraded at elevated temperatures, but they also recover faster after stress is removed [33].

Hot Carrier Injection (HCI), also referred as Channel Hot Carrier (CHC) or Hot Carrier Stress (HCS), is another aging mechanism of CMOS transistors generated by charges that get trapped inside the dielectric. Likewise, its main consequence is observed as the increase in the transistor threshold voltage. The main difference is that HCI is originated from the horizontal electric field between the source and the drain, i.e. when there is a drain current flowing in the channel. HCI is caused by "hot carriers" in the channel that gain enough kinetic energy to enter the dielectric. HCI physics are equivalent to the permanent degradation

of BTI (interface state generation) and it is thus not recoverable [34].

While BTI is impacted by the signal probability, i.e. the time the gate signal spends at a logic value (0 for NBTI and 1 for PBTI), the HCI is impacted by the signal activity, i.e. the average number of transitions. Generally, BTI is more important than HCI in digital circuits because most of the time the signals are in a static state instead of in a transition state [29]. HCI has also an exponential dependence on the supply voltage [35]. However, HCI is not strongly impacted by the temperature unlike BTI [36]. Actually, it even slightly increases at a lower temperature. HCI is therefore as important as, or even more than, BTI at low temperatures.

Besides having different origins, it is incorrect to treat BTI and HCI effects as totally independent mechanisms. This is due to the fact that some defects at the semiconductor-oxide interface are actually shared between both mechanisms. A pessimistic estimation is thus obtained by separately modelling the contribution of each effect and adding them together. However, recent works managed to accurately model the interplay of both BTI and HCI effects [37, 34].

Another issue related to BTI is its induced variability. Transistors of equal size and enduring the same stress do not necessarily produce the same Vth shift. It is due to the amount of preexisting defects inside the dielectric which has a random nature [38]. Past works have already demonstrated that time-zero variability (process-induced) and time-dependent variability (BTI-induced) are not correlated [39]. The local Vth variability therefore increases over time due to BTI. On the other hand, it is reported that HCI has a stronger dependence on time-zero parameters and, therefore, it does not induce more variability [40].

Moreover, the average number of preexisting defects inside the oxide decreases as the technology scales down. This leads to an exponential growth of the time-dependent variability in addition to the time-zero variability increase previously discussed in Section 1.1.2. Figure 1.8 shows the simulated Vth shift for transistors with dimensions equal to $280 \times 720nm^2$ and for transistors with $35 \times 90nm^2$ dimensions [41]. As it can be seen, the increase in variability caused by the reduced number of defects is impressive. Nevertheless, the BTI-induced variability is an important issue for analog circuits and SRAM memories while for digital circuits the mean degradation is more relevant [42].

1.4.2 Destructive aging effects

BTI and HCI are the most important aging effects, but they are not the only ones. While they induce a gradual shift in the transistor parameters, the other effects lead to destructive events which may result in a total failure of the circuit. Time-Dependent Dielectric Breakdown (TDDB) is another phenomenon that occurs inside the dielectric. TDDB is characterized by the creation of a conductive path inside the dielectric caused by trapped charges. Like BTI, the charges are

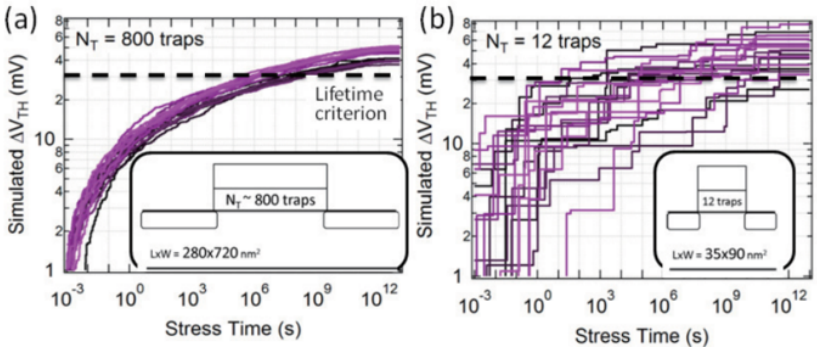


Figure 1.8: Simulated Vth shift due to BTI for transistors with (a) 800 defects and (b) 12 defects. The reduced number of preexisting defects, which is related to transitor size, leads to a considerable increase in local variability [41].

trapped inside the dielectric when a high electric field is applied on the dielectric. This conductive path increases the leakage current and may turn the transistor inoperative. Some soft breakdowns gradually increase the leakage current before the occurrence of the hard breakdown, when a path is created through the gate to substrate.

Electromigration (EM), Stress Migration (SM) and Thermal Cycling (TC) are aging mechanisms that affect the interconnects. Electromigration is due to the momentum transfer between conducting electrons and diffusing metal atoms caused by the current that flows in the wires. EM provokes small voids in the interconnects which may result in an open circuit.

Stress Migration is induced by mechanical stress and can also lead to an open circuit in the interconnects.

Finally, Thermal Cycling is due to elevated temperature gradients. Permanent damage accumulates during large variations of temperature and may eventually lead to failures in the interconnects and packaging.

Models for all the aging effects listed here can be found in [43]. In this thesis we focus only on BTI and HCI effects since we are interested in modeling the circuit performance degradation. The other aging effects are thus not addressed hereafter.

1.5 Conclusion

This chapter introduced all sources of variability in digital CMOS circuits, namely, PVT variations and aging effects. Process variations are caused by manufacturing and atomistic limitations. They can lead to permanent changes in the circuit performance and leakage power from expected values. Voltage variations are mainly due to the circuit impedance. They provoke very fast fluctuations of the

circuit performance, in the order of nano- and microseconds. Local temperature variations are induced by different activities and, as consequence, power dissipation between the circuit blocks. Finally, aging effects, in particular BTI and HCI, manifest through carriers that get trapped inside the dielectric. These aging mechanisms cause gradual degradation of the transistor characteristics reducing the circuit performance over time.

The continuous technology scaling is making hard to design resilient and energy efficient circuits. Reduced transistor size increases both random process variations and aging effects due to the atomistic nature. Each missing/extra dopant or carrier has a greater relative importance as the number of atoms inside a transistor is reduced. Besides, circuits are more susceptible to voltage variations as the voltage scales down while the circuit impedance does not scale at the same rate. Finally, increased transistor density leads to more power density and, consequently, more temperature deviation within a circuit.

The use of safety guard-bands is not anymore suitable due to the considerable waste of energy imposed by it. As shown in Figure 1.9, voltage margins necessary to address the worst case scenario for every source of variability constitute an important part of the supply voltage. The following chapter discusses thus existing solutions to handle variability in digital circuits. These solutions increases the energy efficiency by reducing the safety guard-bands while avoiding timing faults.

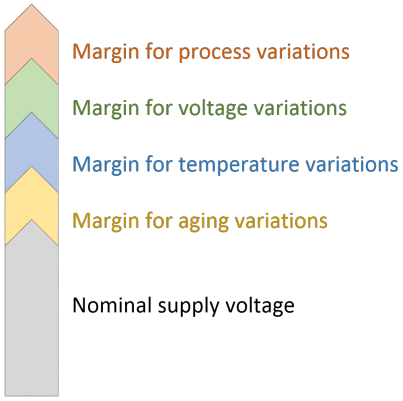


Figure 1.9: Recent technology nodes require large voltage guard-bands to cope with all sources of variability, namely, process, temperature, voltage and aging variations.

Chapter 2

Techniques for coping with variability in digital circuits

In Chapter 1, we introduced the sources of variability and their impact on digital circuits. In the present chapter, we first discuss the traditional approach adopted by designers to handle variability. This approach basically consists in modelling the variations and using simulator tools during the design phase to calculate the timing margins needed to avoid timing faults. However, this approach has become no longer suitable as the technology scaled down and the variability exacerbated because it leads to large guard-bands that do not allow to use the circuit at the best of its capabilities in terms of energy efficiency.

Adaptive techniques have emerged as an important research topic in the last years to reduce the margins fixed by the circuit designers. These techniques implement a strategy to on-line adapt the circuit against variations. The adaptation can be done through a variable supply voltage source or/and a variable clock generator. Some examples are briefly discussed here. We also discuss the adaptive body bias, a technique that has arisen with FD-SOI technologies. Next, we present some sensors for on-line tracking the variability. This comprises monitors for tracking the fluctuation of the critical path timing as well as for directly measuring PVT and aging variations.

Lastly, this chapter analyzes the impact of aging effects on an environmental variability monitor. The use of such monitor is essential in adaptive architectures to track local variations of the supply voltage and the temperature. Still, it is susceptible to aging as any other circuit element. Thus, it becomes less reliable over time. Therefore, we propose a recalibration methodology to mitigate the impact of BTI and HCI on the estimates provided by the monitor making it robust against aging.

2.1 Introduction

Even though asynchronous logic arises as a promising solution to reduce power dissipation and increase robustness to variability [44], nearly all digital circuits are still based on synchronous logic. A synchronous circuit is composed of memory elements, mainly flip-flops, synchronized by a common clock signal. All flip-flops in the circuit simultaneously update their outputs, based on their respective input signals, on the rising edge of the clock (sometimes on the falling edge or even on both edges).

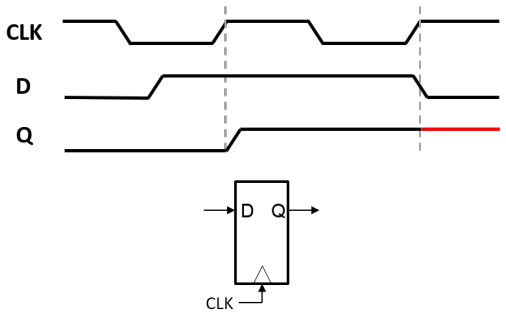


Figure 2.1: Example of a setup time violation occurring in the second clock rising edge. The input data ((D)) arrives at the flip-flop at the same time as the clock rising edge resulting in a timing fault, i.e. the logic value stored in the flip-flop ((Q)) remains '1' instead of changing to '0'.

One of the main concerns for digital circuit designers is thus to avoid setup time violations, also called timing faults. A setup time violation occurs when a data arrives at the flip-flop too close or after the clock rising edge. This makes the flip-flop to store a possibly wrong value. Finally, a timing fault may lead to an error in the circuit operation. This can be catastrophic in safety-critical systems like automotive and aviation ones. Figure 2.1 shows an example of a timing fault. The value latched by the flip-flop in the second clock rising edge is not the correct one. The input data transitions too late, making a logical '1' to be stored in place of a logical '0'.

The data arrival time is determined by the propagation delay of the path that diffuses it. The path propagation delay, in turn, depends on all sources of variability introduced in the previous chapter, namely, process, voltage, temperature and aging. As stated in Session 1.2, the propagation delay is inversely linear to the supply voltage V (for values of V sufficiently higher than the threshold voltage Vth). Therefore, designers usually add voltage margins to V in order to reduce the propagation delay and thus avoid timing faults despite the presence of variations. As shown in Figure 2.2, either a higher voltage (V margin) or a lower frequency (f margin) is adopted.

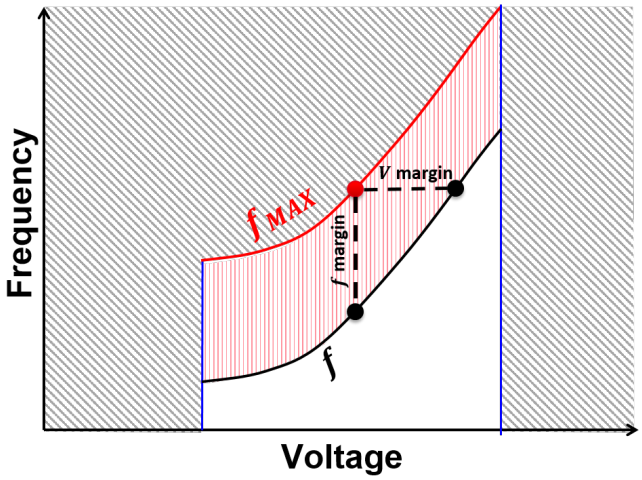


Figure 2.2: Safety margins are used to handle PVT and aging variations. f_{MAX} is the nominal frequency for a given supply voltage, while f is the actual clock frequency taking into account safety margins. These margins can be either a lower frequency (f margin) or a higher voltage (V margin).

Nevertheless, the power dissipated by the circuit strongly depends on V. The power is composed of two components, namely, dynamic and static. Dynamic power is due to the charging and discharging of capacitances originated by the change of state of the transistors. It represents then the power dissipated when the circuit is active. The dynamic power P_{dyn} dominates the power consumption in CMOS circuits and it is related to the supply voltage V as follows:

$$P_{dyn} \propto \alpha f V^2 \tag{2.1}$$

where α is the activity factor, i.e. the percentage of transistors switching, and f is the clock frequency. The static power, on the other hand, is due to the leakage current. It represents the amount of power dissipated when the circuit is in an idle state. Leakage currents depend on Vth and, consequently, on the temperature T. Besides smaller than the dynamic power, the importance of static power has considerably increased with technology scaling due to the reduced Vth. The static power P_{stat} dependence on both V and T can be expressed as:

$$P_{stat} \propto \beta V^{\gamma} e^{\delta T} \tag{2.2}$$

where β, γ and δ are technology- and circuit-dependent parameters. As can be seen from equations (2.1) and (2.2), the use of excessive voltage guard-bands leads to a significant energy loss. Designers must thus accurately estimate the minimum voltage margin needed for a correct operation of the circuit in order to obtain the most energy efficient circuit.

A statistical characterization of the transistor parameters is performed by the semiconductor foundries for each new technology node [45]. This characterization generates device-level models that are used by Electronic Design Automation (EDA) tools to assess the circuit timing characteristics. Besides the mean value, the characterization also provides the standard deviation σ for each transistor parameter. This allows the assessment of the circuit timing for many process variations. For voltage and temperature variations, σ depends on the mission profile, i.e. the circuit application. However, the number of possible corners can reach up to 2^{20} combinations making it impossible to assess the circuit timing for each corner [46]. Therefore, the traditional approach consists in identifying the circuit critical paths and in calculating their propagation delay for the worst-case corner of PVT variations through Static Timing Analysis (STA) [47]. The best- and worst-case values are usually defined as $\pm 3\sigma$ from the mean value.

The main drawback of this approach is that it leads to an over-pessimistic estimation of the circuit performance, the probability of worst-case scenario happening being very low. Actually, in some cases it is even zero due to the correlation between process variations that makes not all conjunction of values feasible. Furthermore, the worst-case process corner is not necessarily equivalent to the worst-case performance corner [48]. Basically, there exist an interaction between transistor parameters so that some specific combinations of values can lead to a performance worse than the combination of worst-case values.

Another way to estimate the safety guard-bands is through Monte-Carlo analysis [49]. It consists in running hundreds or thousands of simulations with random values for the transistor parameters extracted from their respective probabilistic distribution. This is a "brute force" method to statistically evaluate the propagation delay of the critical paths. Its main shortcoming is the excessive computation time required to obtain a representative amount of simulation data. An alternative to Monte-Carlo analysis is Statistical Static Timing Analysis (SSTA) [50]. This approach is similar to STA but it uses probability distributions for the timing of gates and interconnects instead of deterministic values.

To summarize, constant advances in the EDA industry have made the circuit timing estimation a quite matured process. Nowadays, designers have several means for accurately assessing the critical path propagation delay in the worst-case scenario. Nevertheless, the aggressive scaling of the transistor dimensions has exacerbated the impact of variability on digital CMOS circuits, as stated in Chapter 1. Significant voltage margins lead to huge energy losses that cannot be afforded, in particular for mobile applications with limited battery life.

2.2 Adaptive architectures

Adaptive architectures have emerged in the last years to cope with local variability in digital circuits without the need of large guard-bands. Adaptive circuits implement a *sense-and-react* scheme usually based on Dynamic Voltage and Frequency Scaling (DVFS). Basically, a DVFS system integrates two actuators, namely, a variable voltage supplier and a variable clock generator. These actuators dynamically change the circuit supply voltage V and the clock frequency f depending on the required performances. This approach reduces the energy consumption by lowering the power dissipation when there is no need for a high performance. However, in a DVFS system, the V/f couples are predefined in the design phase, with safety margins incorporated to avoid timing faults. The Adaptive Voltage and Frequency Scaling (AVFS) technique consists in employing embedded sensors to track the variability in addition to the actuators used for DVFS. By implementing a closed-loop control, V and/or f can be dynamically changed to adapt against variability. Therefore, AVFS increases further the energy efficiency of the circuit by reducing the safety guard-bands.

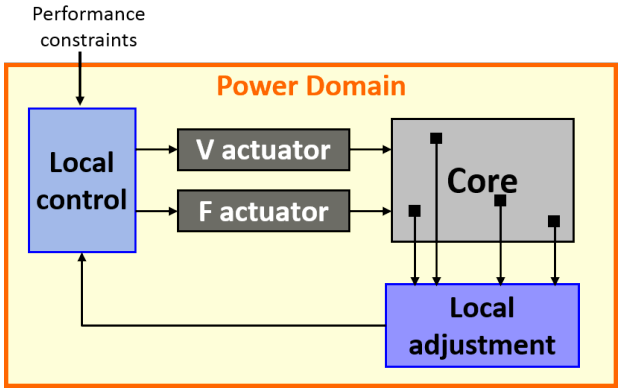
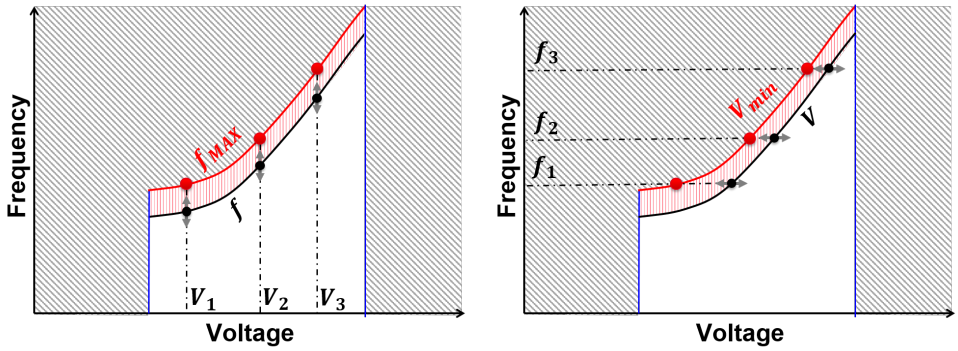


Figure 2.3: General architecture of an AVFS system.

A general architecture of an AVFS system is shown in Figure 2.3. The small black squares inside the *Core* block correspond to the embedded monitors. They provide information regarding the local variability for the *Local adjustment* block. This latter is in charge of checking if the circuit is operating at a safe and energy efficient point. It then notifies the *Local control* block of the need of changing the operating conditions to either avoid timing faults or increase the energy efficiency. Finally, the *Local control* manages both the voltage and the frequency actuators based on this information and on the performance constraints provided by a higher-level control. The objective is to always maintain the clock frequency f as close as possible to the maximum functional frequency f_{MAX}, as shown in Figure 2.2.

2.2.1 Adaptation strategies

In AVFS systems, the change of operating conditions can be done through two variables, the clock frequency and the supply voltage. Generally, while one actuator is used for adaptation purpose, the other remains constant or is set so as to satisfy the performance constraints, like in a DVFS system. A circuit that adapts itself through the clock frequency is called AFS system. As shown in Figure 2.4(a), the values of V are fixed according to the performance levels while f is dynamically changed to deal with variability. Similarly, Adaptive Voltage Scaling (AVS) is when the adaptation is done through the supply voltage, see Figure 2.4(b).



(a) AFS principle, V values are predefined while f is used for adaptation.

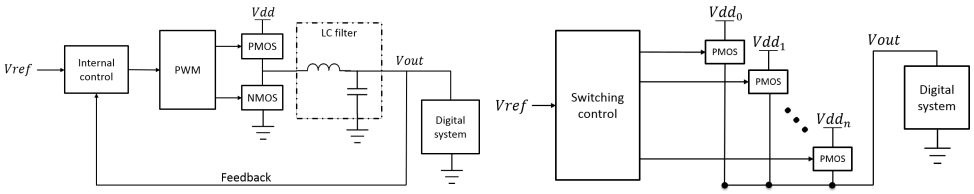
(b) AVS principle, f values are predefined while V is used for adaptation.

Figure 2.4: AFS and AVS strategies with 3 performance levels.

Supply voltage actuator

The first works that focused on adaptive techniques for dealing with static and dynamic variability date back to the 90s [51, 52, 53, 54, 55]. They are all based on AVS which is usually preferred over AFS since the circuit performance is not modified. In these works, the supply voltage V is dynamically changed to its minimum functional value V_{min} based on the information regarding the variability provided by a monitoring system.

In general, a DC/DC buck converter is used as voltage actuator. Figure 2.5(a) illustrates the standard architecture of a DC/DC buck converter. This circuit converts a DC voltage Vdd supplied by the main power supply to a lower DC voltage $Vout$. It is composed of a Pulse Width Modulator (PWM) block that controls the PMOS and NMOS transistors based on an internal ramp. The PWM block is governed by an internal control that generates a command depending on the required voltage $Vref$ and on the current output $Vout$. Finally, an LC filter



(a) Standard architecture of a DC/DC buck converter.

(b) Principle of the Vdd-hopping.

Figure 2.5: Different types of voltage actuators.

reduces the output ripple. This kind of voltage actuator presents great stability and robustness to variability due to its closed-loop control. However, its large area, in particular due to the LC filter, limits its implementation in small circuits. Furthermore, it presents a poor efficiency at low voltages, which is the case of low-power circuits.

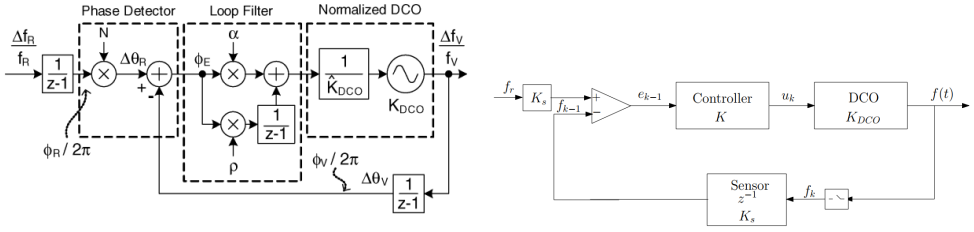
To overcome these limitations, recent works [56, 57, 58] adopted an adaptation strategy based on Vdd-hopping technique, also called voltage dithering. Figure 2.5(b) depicts its principle. Vdd-hopping consists in switching the voltage between two or more predefined Vdd levels to achieve an output voltage V_{out} equals in mean value to V_{ref}. This kind of actuator has a very small area and a fast voltage transition time. However, unlike the DC/DC buck converter, the Vdd-hopping technique has a discrete number of voltage levels. For this reason, Vdd-hopping seems more suitable for changing the performance level (Figure 2.4(a)) than for adapting against variations (Figure 2.4(b)).

Clock frequency actuator

In some cases, AFS may be a better choice than AVS since its implementation is less complex, no impact on the power distribution system, and the time dynamics of frequency actuators are considerably smaller, allowing a faster adaptation to dynamic variations. In an AFS system, the clock frequency f is dynamically controlled to stay always at its maximum functional value f_{MAX}. This is done through a variable clock generator, usually a Phase-Locked Loop (PLL) [59, 60, 61]. PLLs use a Voltage-controlled oscillator (VCO) to generate the oscillating signal. A phase detector compares the output signal frequency with the reference frequency while a filter ensures the actuator stability. Figure 2.6(a) [62] shows an example of a all-digital PLL. This kind of actuator is robust to variability and provides a low-jitter clock signal. However, it has a high area cost which impedes its replication in a multi-core circuit to apply local AVFS in each core.

Another frequency actuator, known as Frequency-Locked Loop (FLL), has thus been developed to overcome this limitation of PLLs. A FLL has a similar

architecture than PLL, also generating a clock signal from a VCO. The difference lies on the way they compare the output frequency with the reference one. In an FLL, the comparison is done directly through the frequency instead the phase. That is why its circuit is much more simpler, as shown in 2.6(b) [63]. The authors in [63] claim that the area of the proposed FLL is 4 to 20 times smaller than a classical PLL. Moreover, its response time is way faster, allowing a transition of frequency level in a few clock cycles. In [58], a digital FLL is used to locally generate a clock signal inside each core of a multi-core circuit with output frequency ranging from 2.9 GHz down to 15 kHz. The area cost was not reported, but the same FLL was adopted in another multi-core architecture [64] with an area overhead of only 0.3% of each cluster.



(a) Example of digital PLL [62].

(b) Example of digital FLL [63].

Figure 2.6: Different types of frequency actuators.

Body bias voltage

Besides AVS and AFS, a new adaptation strategy has recently emerged, Adaptive Body Bias (ABB). MOSFET transistors have a fourth terminal called body which is generally connected to the source. However, it is possible to modulate the transistor threshold voltage Vth by applying a source-to-body voltage, also called Body bias voltage (Vbb). Vth is increased when a positive Vbb is applied, known as reverse body-bias (RBB). This leads to a reduced leakage current with a slower switching speed as side effect. Conversely, forward body bias (FBB), i.e. applying a negative Vbb, reduces Vth and, as a consequence, leads to faster transistors although with increased leakage current.

The use of ABB was first proposed in early 2000s [65, 66, 67]. Initially, it was adopted to compensate die-to-die process-induced Vth variations. For instance, [67] claimed to reduce the frequency variation σ/μ[1] between dies from 4.1% to 0.69% by applying ABB. Better results were achieved by applying different values of Vbb to each circuit block in order to tackle within-die variations. This approach reduced further the frequency variation σ/μ to only 0.21%. The reported area

[1]Where σ is the standard deviation and μ is the average value.

overhead due to the ABB generator and its control block is from 2% to 3% of the total die area.

The development of Ultra-Thin Body and Box Fully-Depleted Silicon-On-Insulator (UTBB FD-SOI) CMOS technologies in the last years has boosted the benefits of ABB. UTBB FD-SOI enables the use of a wide range of body bias voltage. Vbb can vary up to $\pm 3.0V$ in UTBB FD-SOI technologies while in bulk-based technologies it was limited to $\pm 0.5V$ [68]. It is widely reported that FBB increases both the performance and the energy efficiency of digital circuits in UTBB FD-SOI, either at low or high values of Vdd [61, 69]. Many recent works have thus focused on the choice of using dynamic Vdd or dynamic Vbb to find the best performance/energy trade-off. Usually, the best choice is their joint use since each strategy presents better results depending on the operating conditions (clock frequency, temperature, circuit activity, ...) [70, 71].

With regard to reliability, [72] showed that FDSOI and bulk transistors have similar sensitivities to both BTI and HCI effects. Later, [73] demonstrated the benefits of using ABB over AVS in FDSOI technology. Firstly, transistors endure more degradation when a higher Vdd is applied to compensate aging variations instead of using FBB. Secondly, AVS also increases the power dissipation while the power remains constant with ABB. Besides the benefits in energy and reliability, the implementation of ABB is also simpler. Unlike AVS, it does not require substantial consideration of IR drop and it has a low static current since its load is almost purely capacitive [74]. However, Vbb management is not yet a fully mature technique as supply voltage and clock frequency management, this is why AVS and AFS are more popular strategies.

2.3 Monitoring systems

The choice of adaptation strategy and actuators is important to optimize the energy efficiency of the circuit. However, the monitoring system plays a critical role because it is in charge of ensuring that the circuit is operating in a "safe zone", i.e. without timing faults. Moreover, it must ensure that the clock frequency is as close as possible to its maximum functional value. An inaccurate monitoring system can thus result in either an unreliable or an energy inefficient circuit.

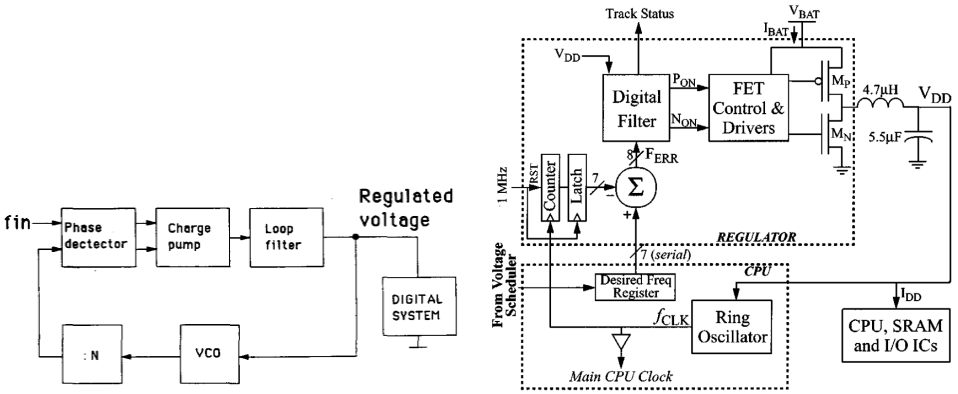
Likewise the adaptation strategy, the monitoring of the different roots of variability in an AVFS system can be done in different ways. They can be split in two major categories. The first one consists in timing fault detection and f_{MAX} tracking techniques. Sensors int this category focus on monitoring the variation of the critical path slack time to detect either the pre-occurrence or the occurrence of a timing a fault. Thus, they use a critical path replica or they are placed on the critical paths. The other category is the direct measurement of the variability. This category of monitors provides a numerical measurement of one

of the source of variations, i.e. process, voltage, temperature or aging.

2.3.1 Timing fault detection and f_{MAX} tracking

The first work to implement a f_{MAX} tracking technique to reduce safety margins was [51]. The authors implemented an AVS system using a PLL to track the variability, as shown in Figure 2.7(a). The supply voltage is regulated until the VCO frequency matches the clock frequency *fin*. This technique is based on the fact that the circuit critical paths and the VCO ring-oscillator have the same sensitivity to PVT variations. This assumption might be true for the technology used in [51] ($2\mu m$), but this is no more valid for the current technologies.

The same concept of variability monitoring has been better developed in [55] for a complete AVFS system. A DC/DC buck converter adjusts its output voltage based on the frequency of a ring oscillator supplied by the DC/DC converter. This ring-oscillator is also used as frequency actuator generating the clock frequency for the CPU, as shown in Figure 2.7(b). Implemented in $0.6\mu m$ technology, the authors claim that a energy reduction of 78% is achieved compared to a system without AVFS. Still, circuits have different sensitivities to static and dynamic variations depending on their topology. Each critical path has a specific topology which is much more complex than the topology of a ring-oscillator. Moreover, such monitor does not endure the same local variability than the logic circuit. Therefore, it is not enough to determine the circuit performance fluctuation in advanced technology nodes.



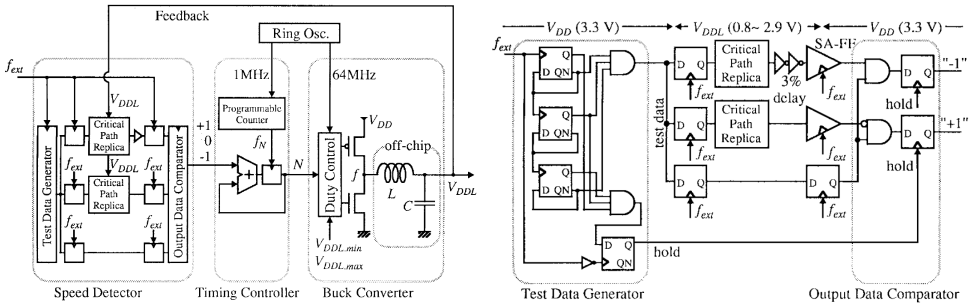
(a) PLL implemented in [51] to adapt the supply voltage against variations.

(b) In [55], the ring-oscillator is used at the same time to generate the clock frequency and to monitor the variability.

Figure 2.7: Pioneer works that used a ring-oscillator to track f_{MAX} [51, 55].

Critical Path Replica

One of the first works to adopt the idea of using Critical Path Replica (CPR) for variability tracking was [54]. Figure 2.8(a) illustrates the proposed AVS system, with a DC/DC converter as voltage actuator and a variability monitor called *Speed Detector*. The *Speed Detector*, shown in Figure 2.8(b), is composed of three paths, each one placed between a pair of flip-flops synchronized by the circuit clock frequency f_{ext}. The first path serves as reference by directly connecting both flip-flops while the second one is a CPR. The last one is a CPR as well but with additional buffers resulting in an increase of the delay of about 3% of the CPR delay. The *Output Data Comparator* compares the output of the three paths. When the output of the third path is equal to the reference one, a "-1" signal is generated and the supply voltage is reduced. When the output of the second path is different from the reference one, a "+1" signal is generated and the supply voltage is increased. The supply voltage remains still only when the reference output is equal to the second output but different from the third one. This technique ensures that the clock period is larger than the critical path delay but no more than 3%. Therefore, it avoids timing faults with a very small timing margin.



(a) AVS system with a DC/DC converter controlled by the output of a variability monitor based on CPR.

(b) *Speed Detector* architecture.

Figure 2.8: AVS system using Critical Path Replica CPR as f_{MAX} tracking technique [54].

Implemented in $0.4\mu m$ technology, the proposed technique improved the performance per Watt by a factor of more than two compared to the previous design [75]. This is achieved with a small area overhead of only 0.5% (without considering the external LC filter of the DC/DC converter). Nevertheless, this technique is limited because it uses only one critical path to track variability. The critical path in modern circuits is not unique and it depends on PVT and aging variations. Various paths must be replicated to obtain an "acceptable" coverage of

the impact of the variability on the circuit. However, it would imply large area and power overheads.

Razor Flip-Flops

Even if it is a perfect replica of the critical path, a canary structure is not able to detect local PVT variability. Furthermore, it may degrade at different rate from the original logic circuit because it does not experience the same local variations. In 2003, [76] conceived the Razor flip-flop, an innovative solution for pipelined processors based on error detection and correction. The technique consisted in adding a shadow latch at the input of the critical paths to detect late transitions, i.e. input data arriving just after the clock rising edge. Later, [77] improved this technique calling it RazorII, as shown in Figure 2.9. RazorII is connected between the master and the slave latches of a flip-flop instead of at its input. Thus, it does not impact the path timing as did the first version of the technique. A programmable delay-chain *DC generator* produces a detection window after the clock rising edge. A transition detector *TDetector* flags an error if a transition occurs at the input within this time window. The incorrect data is then restored in the next clock cycle, resulting in a pipeline stall. Besides timing faults, this technique is also robust to radiation-induced faults known as Single Event Upsets (SEU) [78].

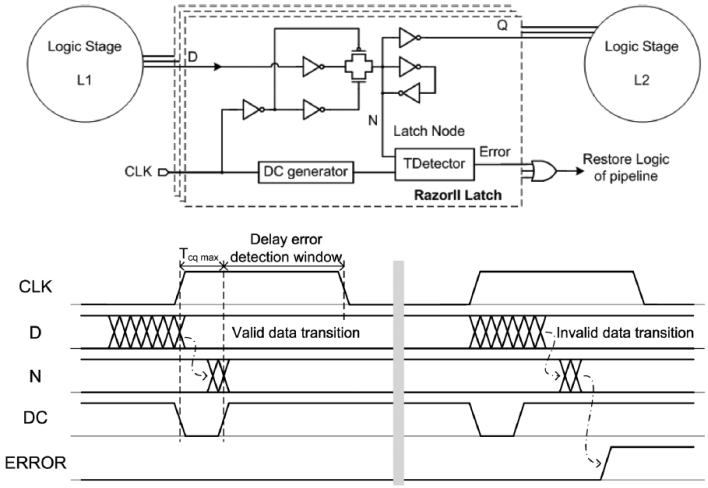


Figure 2.9: Principle of the RazorII Flip-Flop [77]

Unlike CPR, the Razor flip-flop is able to detect local variations since it directly monitors the critical paths. Moreover, it can reach a large coverage with less area overhead because it only inserts a few transistors more at the targeted flip-flops instead of replicating the whole path. Note that several paths are monitored at the same time by instrumenting a single flip-flop. In [77], the RazorII

technique was validated on a 64-bit processor implemented in $0.13\mu m$ technology. In total, 121 out of its 826 (15%) flip-flops were instrumented. The area overhead is not given but the reported power overhead is only 3%. With error detection and correction technique, the energy consumption can be further reduced by lowering Vdd below the Point of First Failure (PoFF), i.e. when the critical path delay becomes larger than the clock period. Actually, the minimum energy point is reached when the energy overhead due to pipeline recovery becomes higher than the energy reduction. On average, 33% energy savings is achieved compared to the worst-case margin approach.

Different implementations of the Razor flip-flop approach have been adopted in many recent works. For instance, [79] developed an architecturally independent version of the technique called Bubble Razor. This version can be automatically inserted into any design without needing detailed knowledge of its internal architecture. The authors employed it in the ARM Cortex-M3 [80], a commercial processor, implemented in $45nm$ CMOS technology. The adaptation can be done either by increasing the clock frequency or reducing the supply voltage. Figure 2.10 shows both the performance and the energy gains for a chip fabricated on the "fast" corner of process variations. As can be seen, the Bubble Razor provides a performance increase of 112% or a energy reduction of 66% compared to a design with safety margins. The authors also implemented a canary structure, CPR, for comparison purpose. The performance and energy gains of the Razor over a CPR are 56% and 42%, respectively.

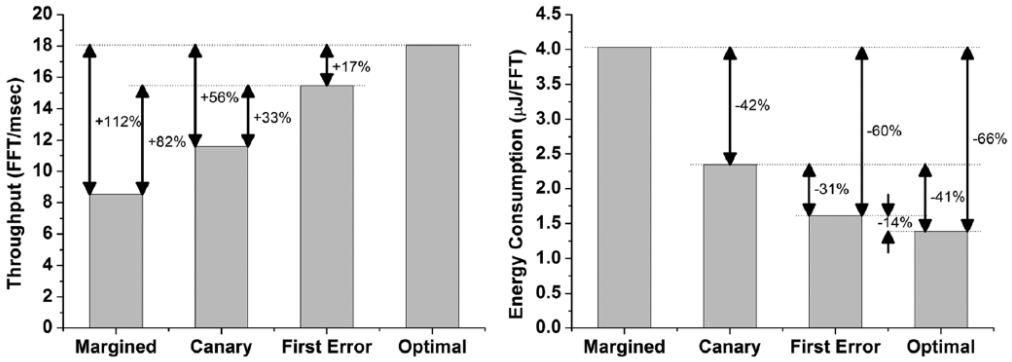
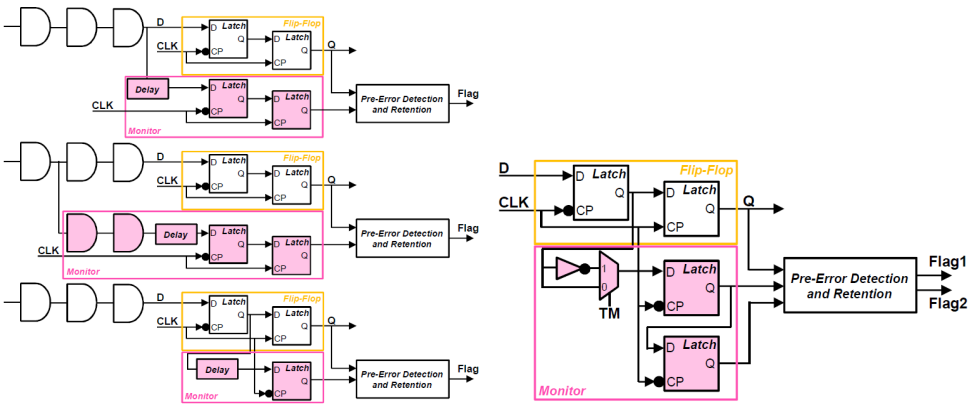


Figure 2.10: Throughput increase and energy reduction achieved with Bubble Razor [79] compared to a margined design and to a CPR approach (canary).

Slack Time Monitoring

Despite the considerable gains achieved with the Razor monitor, its complexity is still a limitation. A large circuitry is needed to allow the pipeline recovery. For example, the Bubble Razor implied an area overhead of 25% [79]. This explains

the emergency of a similar but less complex technique. This technique, usually called In-situ Slack Monitor (ISM), consists in timing fault prediction instead of error detection and correction. The first work to propose such approach was [81]. Similarly to the Razor approach, it is based on shadow flip-flops. The input of the shadow flip-flop is delayed through the insertion of buffers. Therefore, if a late but still valid data transition occurs, it will be latched by the original flip-flop, but not by the shadow one. The two outputs are compared and a warning flag is raised to inform the adaptation system when a timing fault is about to happen. Since the error is detected before it occurs, this technique does not need all the recovery mechanisms required by Razor approaches.

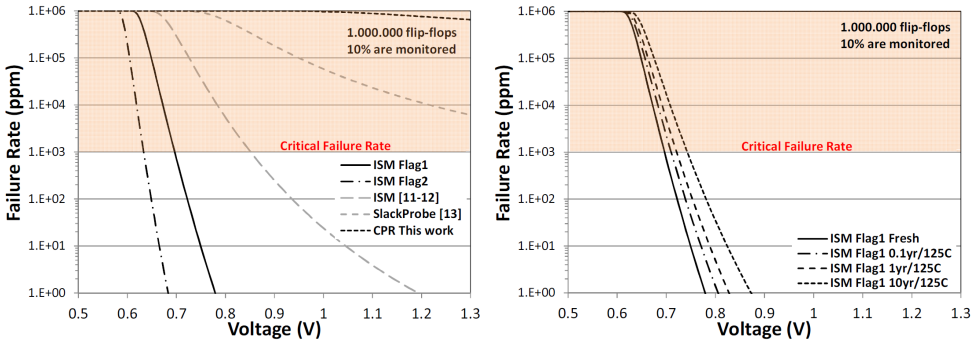


(a) *Top:* Standard approach [81]. *Middle:* SlackProbe [82]. *Bottom:* New approach [83].

(b) Sensor proposed in [83] with two detection windows (Flag1 and Flag2).

Figure 2.11: Different implementations of the In-situ Slack Monitor (ISM) [83].

In [82], the authors suggested the placement of the sensor at an intermediate node of the path instead of at its end in order to increase the monitoring coverage, a technique called SlackProbe. The drawback of this latter approach and the original one is that they impact the path timing by inserting additional loads to it. Moreover, they are not able to detect the local variability that may impact the destination flip-flop. Therefore, [83] proposed a better solution consisting in connecting the sensor between the master and the slave latches. Furthermore, this sensor is composed of only a latch instead of a flip-flop (two latches), resulting in less area overhead. Figure 2.11(a) shows the three implementations. Actually, the sensor in [83] contains a second latch connected to the output of the first one resulting in a second flag, as shown in Figure 2.11(b). Thus, the second flag has a larger detection window than the first one (86 vs 51 under TT, 1V, 25$°C$ conditions). Note that a larger timing window is important to ensure a better pre-error detection at low voltages.



(a) *ISM Flag1* and *Flag2* correspond to [83], *ISM* to [81] and *SlackProbe* to [82].

(b) ISM robustness to aging (small increase in failure rate).

Figure 2.12: Failure rate (parts per million) obtained through Monte-Carlo simulations with local variations [83].

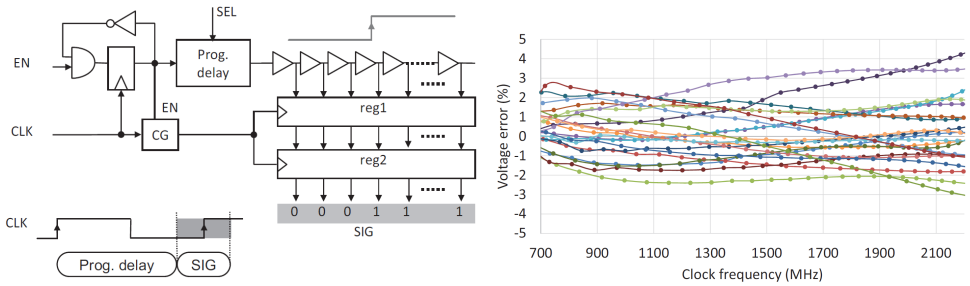
The authors in [83] compared the detection rate of the three solutions and a CPR through Monte-Carlo simulations with local variations. A system with 1 million flip-flops where 10% of them are monitored was considered. The failure rates obtained in parts per million are shown in Figure 2.12(a). The second flag of the sensor in [83] resulted in the smallest failure rate, while the failure rate with CPR is so high that it is hardly seem in the graph. Then, [83] validated the sensor in an AVS system implemented in 28nm FD-SOI technology. 10% of the 869 flip-flops were instrumented, resulting in an area and power overhead of 2.43% and 1.46%, respectively. The total power savings can reach up to 40% depending on the temperature and the clock frequency. Finally, the sensor robustness against aging variations was checked through Monte-Carlo simulations. Figure 2.12(b) shows that, despite the failure rate at low voltages has increased due to aging, it remains within acceptable range.

The main limitation of the in-situ slack monitor, as well as of the Bubble approach, is that it requires the monitored paths being active to sense a variation. For example, a voltage drop can lead to a faulty execution if it occurs right when no monitored path is being excited. Moreover, this approach is based on a detection window with a predefined size. The energy efficiency is sub-optimized if a large window size is chosen, while a small window considerably increases the probability of an undetected timing fault.

A different approach has been recently proposed in [84], called TiMing FaulT (TMFLT) methodology. This technique also uses slack time monitoring, but for calibration purpose instead of on-line monitoring. The calibration is done by overclocking the circuit until functional failure. The objective is to estimate the minimum operating voltage V_{MIN} for each value of the clock frequency. It

is done through a statistical procedure with the warning frequencies of the in-situ monitors. Another sensor based on a delay line is then calibrated with the estimated values for V_{min}. This sensor is shown in Figure 2.13(a). The signature *SIG* generated by the delay line is then used to on-line estimate how close the supply voltage is to V_{min}.

Implemented in a DSP in $28nm$ FD-SOI technology [61], the proposed approach estimates V_{min} with an error within [-2.5%,+3.5%] at nominal clock frequency (1.6GHz). The V_{min} estimation error versus the clock frequency is shown in Figure 2.13(b) for 21 different dies. The authors claim that the proposed technique reduced the DSP power consumption by 19% at the nominal frequency. The main advantage of this approach is that it can detect the pre-occurrence of a timing fault even when the monitored paths are not excited thanks to the delay line based sensor.



(a) Delay line based sensor.

(b) V_{min} estimation error on 21 dies.

Figure 2.13: V_{min} tracking technique based on a calibration phase and a sensor composed of a delay line [84].

2.3.2 Process, voltage, temperature and aging monitors

The monitors introduced in the previous section can be used to detect the impact of variability on the circuit critical paths. Process, voltage, temperature and aging variations lead to fluctuations of the critical path timing. However, these monitors are not able to determine what source of variation has originated the timing fluctuation. A reduction of f_{MAX} may, for example, come from a simple temperature variation or be the consequence of aging degradation. In this section, we present some existing sensors that provide a direct measurement of the variation. This information is important for the local control to decide which adaptation strategy is the most suitable adaptation strategy to mitigate the variation.

Process monitors

As stated in Section 1.1, several transistor parameters may change due to process variations, e.g., threshold voltage, oxide thickness, gate length and width. The deviation of these parameters from their nominal values affects the transistor switching speed and leakage current. Many works have focused on developing sensors to measure these variations after chip fabrication.

For instance, [85] uses temperature measurements to extract the process-induced leakage variation between cores in a multi-core architecture. In [86], a current starved inverter chain is used to characterize the global process corner, i.e. slow, typical or fast. [87] proposes a technique to characterize process variations based on a set of 10 ring-oscillators with different parameter sensitivities. By solving a linear system with the oscillating frequencies and a predefined sensitivity matrix, it is then possible to estimate the variation of the gate length and of both PMOS and NMOS Vth. A similar sensor is presented in [88] but with reduced area. The process corners of PMOS and NMOS transistors are estimated by using only two ring-oscillators, one of them being more sensitive to variations in PMOS while the other is more sensitive to variations in NMOS. Finally, [89] implements an array-based sensor with 256 test units to characterize local variations of drain current and Vth.

Voltage and temperature monitors

In general, the most efficient voltage and temperature monitors are analog circuits. For instance, the sensor presented in [90] estimates the temperature within the chip with a standard deviation σ of $\pm 0.05°C$ over a wide operating range from $-55°C$ up to $125°C$. [91] and [92] designed sensors to detect fast voltage drops, while [93] proposed a monitor to measure supply voltage noise with a voltage and time resolution of 5mV and 0.4ns, respectively. The main drawback of these monitors is that, as they are analog circuits, they need an Analog-to-Digital Converter (ADC) to generate a digital output. As this implies a large silicon area overhead, the replication of these sensors is very costly in terms of area. They are thus not suitable for monitoring local variations.

Other works focused on designing digital monitors to reduce the required area. For instance, [94] proposed a small temperature monitor based on delay lines having a resolution of $0.16°C$ and measurement errors within $\pm 0.9°C$. However, it has been fabricated in an old technology ($0.35\mu m$) which is less sensitive to variability. Some works also proposed small area temperature monitors based on a single ring-oscillator [95, 96, 97], but they all share the same shortcoming than [94]. Moreover, [98] demonstrated the limitations of using this kind of sensors to measure temperature in low-voltage circuits due to the increased sensitive to voltage variations.

The main difficulty in measuring the temperature from the oscillating frequency of a ring-oscillator is that it also depends on other sources of variation, in particular the voltage. One solution consists in using two or more ring-oscillators to isolate the temperature and the voltage effects. [99] proposed a thermal sensor composed of two differential ring-oscillators. In this way, the sensor is robust to process and voltage variations. Similarly, [100] used a small sensor composed of 7 ring-oscillators to track dynamic variations. A data fusion technique is applied on the oscillator frequencies to estimate both the voltage and the temperature. Implemented in $32nm$ technology, it has a reduced area of $450\mu m^2$. It is capable of providing voltage and temperature estimates with errors below 5mV and $7°C$, respectively. A calibration method is executed after circuit fabrication to take process variations into account.

Even tough the voltage and temperature monitor proposed in [100] is robust to process variations, it is still vulnerable to aging variations. In other words, it is not guaranteed that the sensor will continue work properly after some time of use. [101] developed a similar sensor to estimate both the voltage and the temperature. The authors claim that the architecture they adopted for the ring-oscillators make their sensor robust against NBTI variations. However, it has not been demonstrated through any aging experiment. Furthermore, the sensor is still vulnerable to HCI.

Section 2.4 will address the issue of developing an aging-robust sensor for dynamic variability.

Aging monitors

Aging monitors were widely proposed in the last years since aging has become a major issue. Most of them focus on characterizing both N/PBTI and HCI effects. They are based on a dedicated circuit composed of two identical ring-oscillators. Their principle is to degrade one of the ring-oscillators while the other one is not powered and, consequently, not stressed. It is then possible to measure the aging degradation by measuring the beat frequency, i.e. the difference between the two ring-oscillator frequencies.

For instance, [102] proposed a sensor for separately measuring the frequency degradation due to BTI and HCI. It comprises two pairs of identical ring-oscillators (ROSC), as shown in Figure 2.14. The first ROSC, called *BTI_ROSC*, is kept in a steady mode. This ROSC endures only BTI degradation, since HCI occurs during the state change of transistors. Meanwhile, the second ROSC, *DRIVE_ROSC*, is always in oscillating mode thus being degraded by both BTI and HCI effects. The frequency shift of the first ROSC gives the BTI degradation. By subtracting it from the frequency shift of the second ROSC, we then have the HCI degradation. Note that this approach is valid only if both BTI and HCI effects are supposed additive (linear effects).

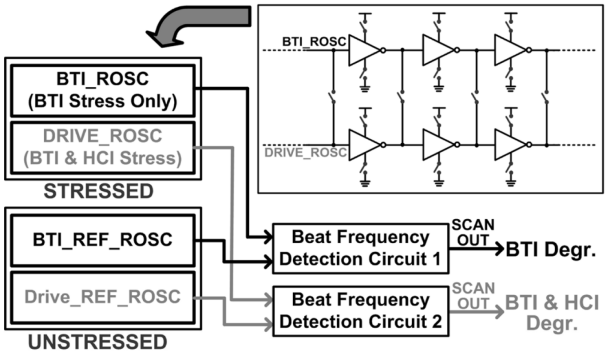


Figure 2.14: Ring-oscillator (ROSC) based sensor for separately measuring BTI and HCI effects [102].

In [103], a very small NBTI monitor is proposed. It relies on a PMOS transistor that is used to starve the current supplied to a ring-oscillator. Since the NBTI-induced Vth degradation reduces the transistor drain current, the ring-oscillator frequency will shift with it. The authors claim that a Vth shift of 10% leads to a 53% change in the oscillator frequency. The sensors are arranged in an array forming a bank. The bank also includes a counter and three registers which allow four quick measurements. Each bank contains 16 sensors and each die has 6 banks, resulting in 96 NBTI sensors in total. Such array-based sensor allows fast characterization of the statistical nature of NBTI. Manufactured in $130nm$ technology, the whole monitor has a small silicon area of $308\mu m^2$, 110 times smaller than previous similar work [104] in same technology.

Many other works also proposed ring-oscillator based aging sensors [105, 106, 107, 108, 109, 110, 111]. Besides having different architectures, they all consist in measuring the frequency beat of two identical structures. The main drawback of this kind of monitor is that it measures the degradation of a dedicated structure instead of the logic circuit itself. The circuit topology has an important role on the aging degradation. Thus, it is not possible to assume that the ring-oscillator and the datapath experience the same degradation.

Some works [105, 108] aimed to tackle this limitation by using critical path replicas as ring-oscillators. Nevertheless, the actual and the replica datapath can endure different stresses due to local PVT variations. Furthermore, numerous monitors would be needed to cover all the possible critical paths, leading to a considerable area overhead. Lastly, but most importantly, both BTI and HCI effects have a strong dependence on the workload which is difficult, or even impossibly, to reproduce on a canary structure. Figure 2.15 [112] illustrates the limitation of using a CPR to measure the aging degradation. As can be seen, the actual datapath (f_{MAX}) presents a considerably wider spread of degradation than the CPR. It is due to the workload dependence and the local variations.

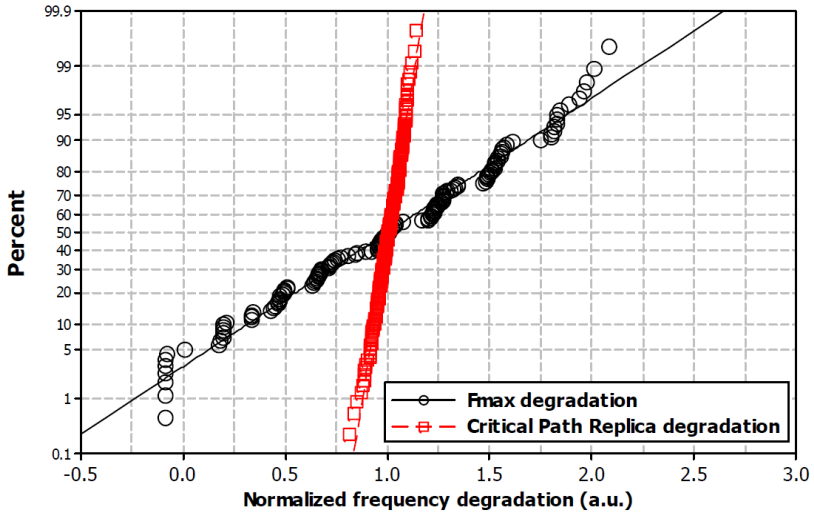


Figure 2.15: Distribution of the normalized degradation of the logic circuit (black circles) and of the CPR (red squares) [112].

2.4 Aging mitigation of a VT monitor

Section 2.3.2 has presented some of the state-of-the-art sensors to directly track voltage and temperature variations. Still, these sensors are affected by process and aging variations as any other circuit element. Most of them do consider the impact of manufacturing mismatches. They either design a process-tolerant sensor or implement a process-aware calibration method that is conducted after the circuit fabrication. However, aging effects are still a recent concern and they are not addressed by these works. The only exception seems to be [101] where the authors claim to adopt an NBTI-resilient architecture for their sensor. Nevertheless, its resilience has not been demonstrated through any aging test or simulation.

As stated in Section 1.4, the most important aging effects in CMOS circuits are the BTI and the HCI [29]. Besides impacting the performance of the processing elements, these mechanisms also affect the information provided by the integrated sensors. For instance, [113] developed a low-cost small-size digital probe made of 7 ring-oscillators (ROs). This so-called Multiprobe, together with a data fusion technique [100], allows to estimate the V and T values within the chip. Nevertheless, [100, 114] do not consider the influence of aging on the accuracy of the V and T estimates.

Therefore, in this section we firstly summarize the Multiprobe sensor and its VT estimation method. Then, we analyze the efficiency of the Multiprobe when its ROs are degraded by both BTI and HCI effects. Finally, we propose a recalibration method to guarantee the sensor accuracy despite aging effects. The results presented in this section have been published in [9].

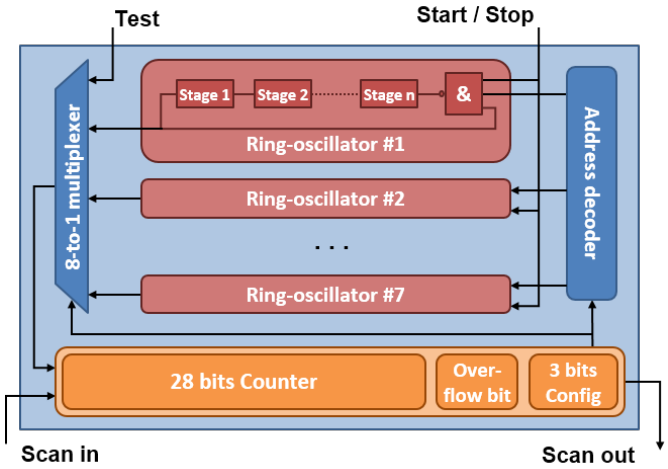


Figure 2.16: Multiprobe sensor for dynamic variability monitoring [114].

Table 2.1: Description of the 7 ROs composing the Multiprobe.

Name	Delay cell	Stages	F_{RO}^{nom}
NCap	2 inverters loaded with capacitors made of a NMOS transistor	6	1.53 GHz
PCap	2 inverters loaded with capacitors made of a PMOS transistor	6	1.46 GHz
XOR	1 std cell XOR	8	1.75 GHz
Inverter	2 std cell inverters	13	1.49 GHz
Latch	1 std cell latch	5	1.72 GHz
LongWire	2 inverters linked together with long wires using several metal layers and vias	13	1.69 GHz
LowTherm	2 inverters specially conceived to be more sensible to the temperature	2	2.29 GHz

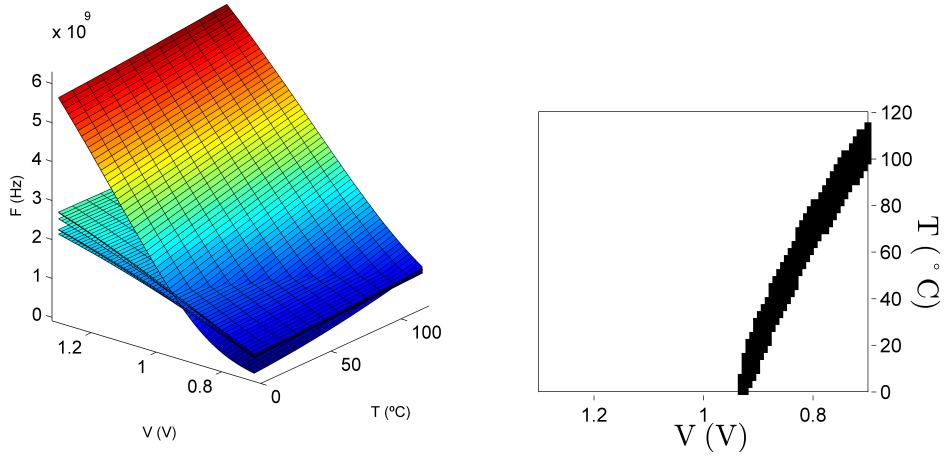
2.4.1 Multiprobe: an all-digital on-chip sensor to monitor VT variations

The Multiprobe is an all-digital sensor designed only with standard cells to ease its integration in an MPSoC design. Seven ROs and a digital counter constitute its main blocks, as can be seen in Figure 2.16. Neither costly analog blocks nor Analog-to-Digital Converter (ADC) are thus needed. Each RO is composed of several delay cells and a NAND standard cell. Table 2.1 gives detailed information about each RO, where F_{RO}^{nom} is the respective oscillating frequency in $28nm$ technology for the nominal condition $(P, V, T) = (TT, 1V, 20\,^{\circ}C)$.

The ROs were purposely designed with different delay cells so as to exhibit distinct sensitivities to VT variations. The *LowTherm* RO is made of special current starved inverters which enhance its temperature sensitivity. This feature

makes the estimation of both the voltage and the temperature possible taking also into account the oscillating frequency of other ROs. Figure 2.17(a) shows the frequency surfaces of all the seven ROs on the $\{V, T\}$ plan. The Mutiprobe has a reduced silicon area of $450\mu m^2$ in $28nm$ technology. All ROs are therefore supposed to experience the same PVT-variability (static and dynamic). Due to its small-size, the Multiprobe can be duplicated within the chip, which is highly suitable for fine-grain AVFS architectures.

The ROs are *per se* multi-sensitive sensors. As a consequence, an appropriate data fusion technique must be used to estimate the V and T values the Multiprobe is experiencing. For such, a database with all ROs' frequencies for several $\{V, T\}$ conditions over the operating ranges is necessary. This database is constructed through post-layout simulations and it is stored in an external memory. Since process variations highly impact the ROs' frequencies, a calibration method has been proposed in [115]. The database is initially constructed considering the Typical-Typical (TT) corner. The actual process corner of the Multiprobe is characterized at the chip start-up. A correction ratio is then computed and applied to all models in the database. The detailed description of this calibration method is given in [115].



(a) Frequency surfaces of all 7 ROs.

(b) $\{V, T\}$ conditions where the frequencies of all ROs are very close ($< 10\%$).

Figure 2.17: Frequency surfaces of all 7 ROs. Note that there are some $\{V, T\}$ conditions where all ROs have very similar frequencies. In these conditions the estimation method is less accurate.

Voltage and temperature estimation method

The frequencies of all ROs have to be firstly quantified in order to perform a VT estimation. The measure is done through the integrated counter, one RO at a time. Both V and T are then estimated from the seven oscillating frequencies using a set of Kolmogorov-Smirnov (KS) goodness-of-fit tests [100] which are non-parametric hypothesis tests. For two empirical samples of size n, the KS test estimates if both samples come from the same distribution law. The estimation method is depicted on Figure 2.18.

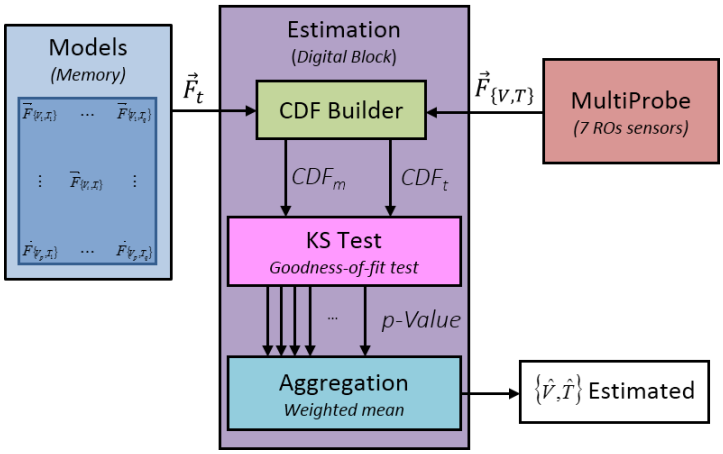


Figure 2.18: VT estimation method principle.

The *CDF Builder* block computes for each measurement vector $\vec{F}_{\{V,T\}}$ a Cumulative Distributive Function (CDF). The CDF is computed from the sums of all pairs of frequencies. This gives a richer representation of the sensor behavior under the current $\{V,T\}$ condition. A CDF_m is computed from the Multiprobe measurements $\vec{F}_{\{V,T\}}$ and another CDF_t is computed from the $\vec{F}_t$ stored in the *Models Database*. As stated above, the database is constructed through post-layout simulations from measurements $\vec{F}_{\{V_i,T_j\}}$ corresponding to the condition $\{V_i,T_j\}$. A calibration is later performed on the fabricated chip to take into account the process variation [115]. One KS test evaluates if the CDF_m and the CDF_t are similar and thus correspond to the same $\{V_i,T_j\}$ conditions. In the *KS Test* block, the maximum gap between both Cumulative Distribution Functions CDF_m and CDF_t is first computed:

$$D_t = \sup_x |CDF_m(x) - CDF_t(x)| \tag{2.3}$$

Then, the probability p_t (called p-Value) that CDF_t and CDF_m come from the same distribution is given by:

$$p_t(\lambda) = 2\sum_{k=1}^{+\infty}(-1)^{k+1}e^{-2k^2\lambda^2} \text{ with } \lambda = \sqrt{n}\cdot D_t \tag{2.4}$$

Lastly, the *Estimation* block computes the estimated $\hat{V}$ and $\hat{T}$ values. For each KS test, the p_t value is collected in $\vec{P} \in \mathbb{R}^M$. Then, s CDFs that "best" match CDF_m (via the comparison of p_t to a given threshold) are used in the *Aggregation* block to compute $\{\hat{V}, \hat{T}\}$, using $\{V_i, T_j\}$ associated with the s CDFs. In the present work, a weighted mean is used:

$$\hat{V} = \frac{\sum_{k=1}^{s}(p_{t_k} \cdot V_k)}{\sum_{k=1}^{s} p_{t_k}} \ , \ \hat{T} = \frac{\sum_{k=1}^{s}(p_{t_k} \cdot T_k)}{\sum_{k=1}^{s} p_{t_k}} \tag{2.5}$$

The validation of the estimation method has been described in details in [100]. Its accuracy highly depends on the number of models stored in the database. The same goes for the memory overhead, the complexity of the calibration phase and the maximum estimation throughput. A deep analysis of the choice of the database size is given in [100]. The estimation method can be implemented in software, but it would be then impossible to monitor the dynamic variations within an adequate interval. Therefore, a hardware accelerator with a complexity equivalent to 9 kgates has been designed to execute the estimation method. The accelerator can compute a new $\{\hat{V}, \hat{T}\}$ within $25\mu s$ at 600MHz for a database containing 366 models [115].

Here, the VT estimation method is evaluated through Matlab. The model database is constructed by ranging V from $0.7V$ to $1.3V$ with a step $\Delta V = 10mV$ and T from $0\,^\circ C$ to $120\,^\circ C$ with a step $\Delta T = 10\,^\circ C$. The final database contains 793 $\{V, T\}$ models at total. A broader range of $\{V, T\}$ conditions is considered to validate the proposed VT estimation approach. For such, 3567 different $\{V, T\}$ points are tested. The same ranges are conserved, but with steps $\Delta V = 7mV$ and $\Delta T = 3\,^\circ C$. The mean absolute estimation errors for V and T are respectively:

$$\mu_{|\epsilon_V|} = 2.9mV, \quad \mu_{|\epsilon_T|} = 6.07\,^\circ C \tag{2.6}$$

while the mean signed estimation errors are close to zero:

$$\mu_{\epsilon_V} = 0.18mV, \quad \mu_{\epsilon_T} = -0.79\,^\circ C \tag{2.7}$$

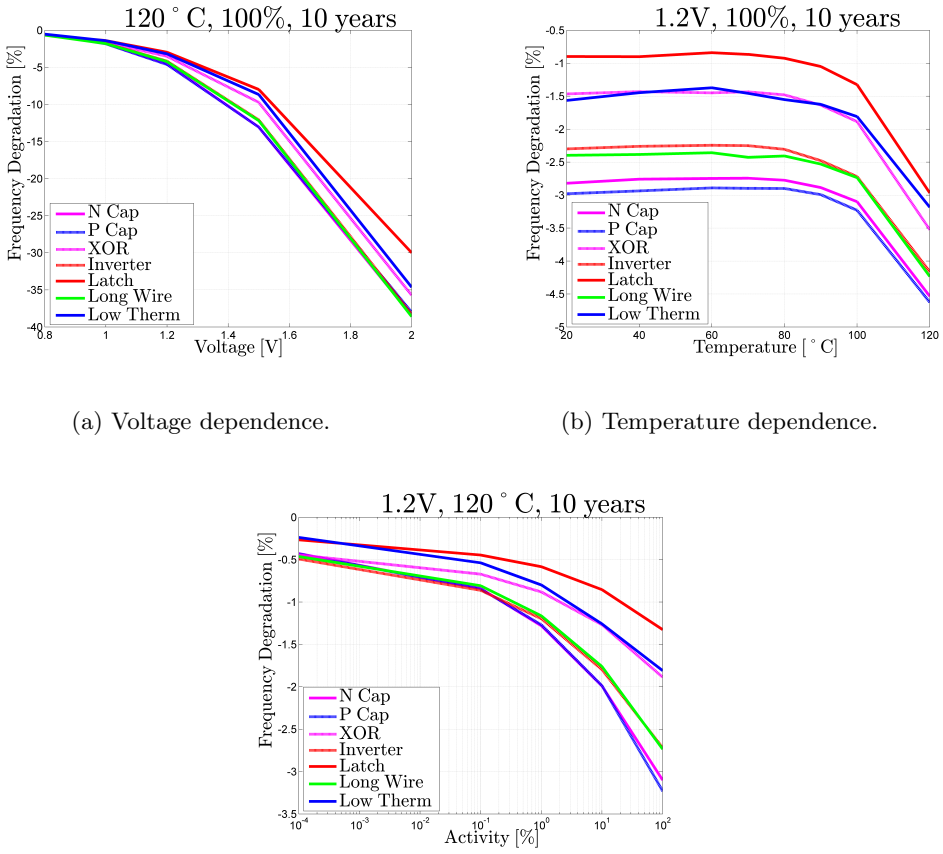
with standard deviations equal to:

$$\sigma_{\epsilon_V} = 5.3mV, \quad \sigma_{\epsilon_T} = 9.29\,^\circ C \tag{2.8}$$

2.4.2 Impact of BTI and HCI effects on measurements

The Multiprobe and its VT estimation method were already implemented and validated on a prototype circuit in $32nm$ technology [58]. However, its reliability has never been studied so far. As any other integrated circuit, it is vulnerable to aging degradation. Since these variations lead to parametric shifts in the circuit, the frequencies of all ROs are supposed to shift over time.

Therefore, to assess the reliability of the Multiprobe, SPICE simulations with aging variations are conducted on the post-layout netlists of all ring oscillators implemented in 28nm FD-SOI technology. The simulations are performed through the Eldo User-Defined Reliability Model (UDRM) API [116]. This API computes the stress experienced by each transistor during a transient simulation and then runs a new simulation taking the degradation into account. A state-of-the-art model coupling both BTI and HCI effects is adopted here [112, 34]. As stated in Section 1.4.1, both effects increase the threshold voltage of the transistors, which results in an increase in the gate delay and, consequently, a reduction of the ROs' frequencies. More details about device-level aging simulation will be later given in Section 3.3.2.



(a) Voltage dependence.

(b) Temperature dependence.

(c) Activity dependence.

Figure 2.19: Aging dependence on (a) Voltage, (b) Temperature and (c) Activity (toggle rate).

Three main factors determine the final degradation experienced by each RO, namely, the supply voltage (V), the temperature (T) and the activity (A), the

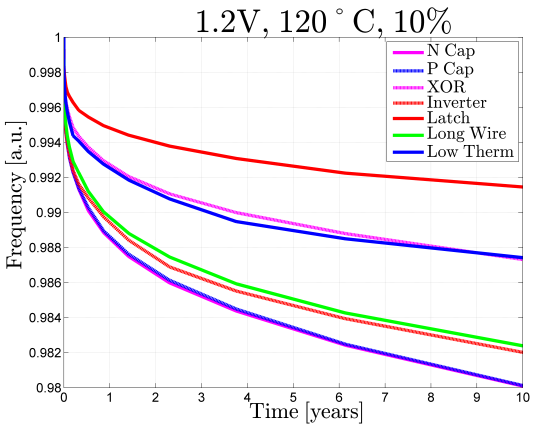


Figure 2.20: Frequency evolution of the seven ROs for 10 years of aging. Stress conditions $(V, T, A) = (1.2V, 100\,^{\circ}C, 10\%)$.

last one being the percentage of time the RO is oscillating. Figure 2.19 shows the resulting frequency degradation of all ROs after 10 years for different values of V, T and A. The degradation has an exponential dependence on both the supply voltage and the activity. Nonetheless, it is exponential dependent on the temperature only for values above $80\,^{\circ}C$. There is no temperature dependence observed for lower temperatures. This is due to the fact that the predominant source of degradation at high temperature is the BTI, which has an exponential dependence on the temperature. Meanwhile, HCI becomes more dominant at low temperature, thus canceling the temperature dependence.

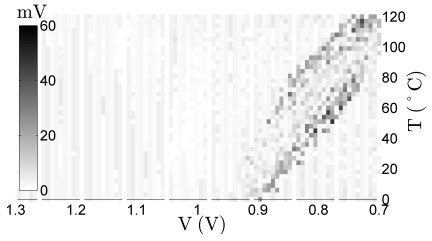
The stress conditions adopted for voltage, temperature and activity are $1.2V$, $100\,^{\circ}C$ and 10%, respectively. The operating conditions will never be constant in a real application as they evolve during the circuit lifetime. Nevertheless, these values represent a worst case scenario. Figure 2.20 shows the aging behavior of all ROs for up to 10 years of stress. The worst frequency degradation observed is almost 2% for both NCap and PCap ROs. Note that their degradation reaches more than half of it after only 1 year of stress. Indeed, due to its exponential nature, the aging is much more important during the first months of the circuit life. Moreover, the frequency shift is not the same for all $\{V, T\}$ conditions. Since the RO frequency is proportional to the difference between the supply voltage and the threshold voltage, it is more sensitive to aging variations at lower values of V. Figure 2.22(a) shows the relative frequency shift of the NCap RO over the whole $\{V, T\}$ plan. The resulting shift is 3 times more important when V is near $0.7V$ than when V is higher than $1V$.

Table 2.2 shows the estimation errors obtained for a fresh Multiprobe as well as for 3 aged conditions, namely, 1, 3 and 10 years of stress. The mean absolute (signed) estimation errors for voltage and temperature are represented

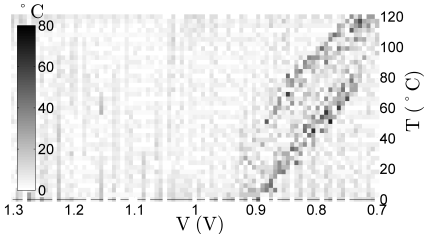
Table 2.2: Mean (absolute and signed) estimation errors and standard deviations for different aging situations ($V = 1.2V, T = 100\,^\circ C, A = 10\%$)

	Fresh	1 year	3 years	10 years
$\mu_{\lvert\epsilon_V\rvert}$	$2.9mV$	$7.9mV$	$9.8mV$	$12.7mV$
$\mu_{\lvert\epsilon_T\rvert}$	$6.07\,^\circ C$	$7.62\,^\circ C$	$8.37\,^\circ C$	$9.80\,^\circ C$
μ_{ϵ_V}	$0.18mV$	$7.2mV$	$9.3mV$	$12.3mV$
μ_{ϵ_T}	$-0.79\,^\circ C$	$-4.38\,^\circ C$	$-5.60\,^\circ C$	$-7.06\,^\circ C$
σ_{ϵ_V}	$5.3mV$	$5.7mV$	$5.5mV$	$5.8mV$
σ_{ϵ_T}	$9.29\,^\circ C$	$10.16\,^\circ C$	$12.02\,^\circ C$	$10.71\,^\circ C$

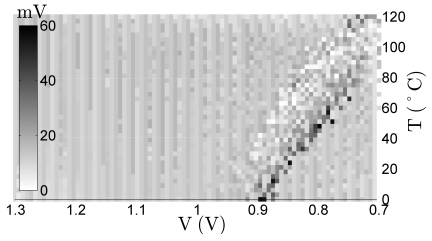
by $\mu_{|\epsilon_V|}$ (μ_{ϵ_V}) and $\mu_{|\epsilon_T|}$ (μ_{ϵ_T}), while σ_{ϵ_V} and σ_{ϵ_T} stand for the associated standard deviations. The error is calculated as the difference between the real value and the estimated one. The increase in the mean error is thus explained by the frequency degradation of all ROs. Since the frequencies are now lower than before, the estimated voltage is also lower than the correct one. This offset can be noted through the mean voltage estimation error. For a fresh sensor, it is equal to $0.18mV$ and it raises to $12.7mV$ after 10 years.



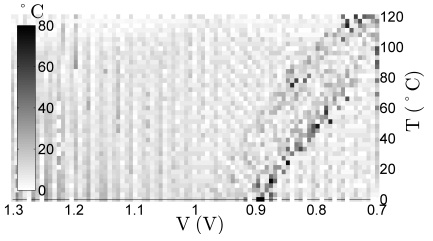
(a) V estimation error map for a fresh Multiprobe.



(b) T estimation error map for a fresh Multiprobe.



(c) V estimation error map for a stressed Multiprobe.



(d) T estimation error map for a stressed Multiprobe.

Figure 2.21: Maps of voltage and temperature absolute estimation errors on the whole $\{V, T\}$ plan. Above, for the simulation of a fresh Multiprobe. Below, for a simulation considering 10 years of stress.

Figures 2.21(a) and 2.21(b) show the distributions of the absolute errors on the estimation of V and T, respectively, over the plane $\{V, T\}$ for a fresh Multiprobe. The errors are very small for most of the $\{V, T\}$ conditions. However, they highly increase in the diagonal of the cartography. Actually, all ROs have almost the same oscillating frequency in this area because of their design properties. It is therefore very hard to extract the V and T values from the measurements using the KS test. Fig. 2.17(a) shows the frequency surfaces for the seven ROs of the Multiprobe. Fig. 2.17(b) highlights the $\{V, T\}$ conditions where all ROs have nearly the same oscillating frequencies. Actually, the dark area corresponds to the operating points $\{V, T\}$ where the maximal absolute difference between two frequencies is below 10%. Figures 2.21(c) and 2.21(d) illustrate the distribution error for a Multiprobe after 10 years of aging. The larger errors are placed in the same positions of the cartography, see Figures 2.21(a) and 2.21(b). However, the errors in the rest of the plane become relevant too. This explains the increase of the mean absolute estimation error observed in Table 2.2.

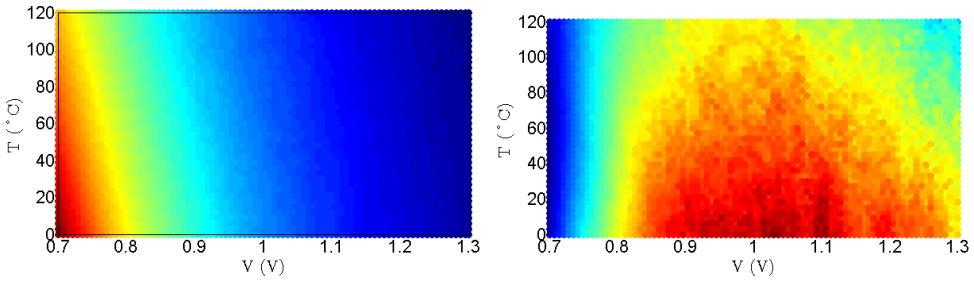
2.4.3 Aging-aware recalibration proposal

As shown in the previous section, aging variations on the Multiprobe lead to an important deviation of the VT estimated values. After just one year of operation, the voltage mean absolute error increased from $2.9mV$ to $7.9mV$, more than 170%. Thus, if no aging-aware technique is applied in its design or execution, the VT estimation becomes inaccurate after just a few months of circuit life.

Some models allow the prediction of the threshold voltage (V_{th}) shift due to aging. However, the modelling of the aging variation is not an easy process since the resulting degradation is strongly dependant on the stress conditions (V, T, A). Even if it was possible to track on-line the operating conditions of the circuit, a precise aging model would require a complex computation on a large dataset. Each RO is composed of several logic gates which in turn are composed of many transistors. The V_{th} shift of each transistor would have to be computed to reach the final RO frequency degradation.

An advantage of the Multiprobe over similar sensors is that its VT estimation method is based on a database. It can thus be reprogrammed when needed, considering that its database is stored in a flash memory. Basically, the recalibration of the Multiprobe consists in reconfiguring its database to cope with aging variations. The question is then how to recompute the database models.

The ideal case would be to measure again the seven ROs frequencies for all the $\{V, T\}$ conditions. The inconvenience is that it is not possible to change both the temperature and the voltage with such required precision in a real-life environment. Thus, a simplified method is to measure the degradation in a known $\{V, T\}$ condition and apply it to all models. Even though the aging variation is not the same over the whole $\{V, T\}$ plan (see Figure 2.22(a)), this method can



(a) Relative degradation (%). (b) Absolute degradation (Hz).

Figure 2.22: Map of the frequency shift of the NCap RO after 10 years of stress, with the lowest degradation values in blue and the highest ones in red.

decrease the offset between the oscillating frequencies of the aged ROs and those stored in the database.

Two approaches can be applied to perform this correction. The first one consists in applying a correction factor based in the relative degradation (i.e. in percentage), as shown in Figure 2.22(a). This correction factor is obtained with:

$$r_{rel} = F^{aged}_{(Vp,Tp)} / F^{fresh}_{(Vp,Tp)} \tag{2.9}$$

where (Vp, Tp) is a point whose both temperature and voltage are known, $F^{fresh}_{(Vp,Tp)}$ is the frequency stored in the database for this point while $F^{aged}_{(Vp,Tp)}$ is the new frequency measured. This correction factor is calculated for each one of the seven ROs. Then it is applied to each (Vi, Tj) condition in the models database:

$$F'_{(Vi,Tj)} = F_{(Vi,Tj)} * r_{rel} \tag{2.10}$$

Another way of doing this recalibration is using the absolute amount of frequency decrease as a correction factor (i.e. in Hertz). Figure 2.22(b) shows the absolute degradation of the NCap RO after 10 years over the whole $\{V, T\}$ plan. The correction factor in this case is calculated as:

$$r_{abs} = F^{aged}_{(Vp,Tp)} - F^{fresh}_{(Vp,Tp)} \tag{2.11}$$

then, it is added to all models in the database:

$$F'_{(Vi,Tj)} = F_{(Vi,Tj)} + r_{abs} \tag{2.12}$$

The second recalibration method is simpler than the first one because it requires only addition and subtraction. Regarding their complexity, these algebraic operations are considerably easier to be executed than multiplication and division. Moreover, this method is also theoretically better since the difference between the

Table 2.3: Comparison of estimation errors for different database recalibration methods after 10 years of aging

	No recalibration	Ideal case	Absolute correction	Relative correction
$\mu_{\lvert\epsilon_V\rvert}$	$12.7mV$	$3.1mV$	$3.3mV$	$4.7mV$
$\mu_{\lvert\epsilon_T\rvert}$	$9.80\,^\circ C$	$6.80\,^\circ C$	$7.13\,^\circ C$	$7.43\,^\circ C$
μ_{ϵ_V}	$12.3mV$	$-0.12mV$	$-0.15mV$	$-0.20mV$
μ_{ϵ_T}	$-7.06\,^\circ C$	$-0.45\,^\circ C$	$-0.98\,^\circ C$	$-0.25\,^\circ C$
σ_{ϵ_V}	$5.8mV$	$6.2mV$	$6.4mV$	$7.1mV$
σ_{ϵ_T}	$10.71\,^\circ C$	$11.21\,^\circ C$	$11.40\,^\circ C$	$11.85\,^\circ C$

maximum and minimum values for the absolute degradation is smaller than for the relative degradation.

We then analyzed the VT estimation errors after 10 years of aging, using both calibration approaches. The (Vp, Tp) point chosen is equal to $(1V, 60\,^\circ C)$, which is the center point of the $\{V, T\}$ plan. Table 2.3 shows the mean estimation errors and standard deviations obtained for both calibration methods. The first column contains the results using the original database, without recalibration. In the second column the results for the ideal case are provided, where all ROs frequencies were remeasured for each one of the $\{V, T\}$ conditions. The two last columns stand for the recalibration using an absolute and a relative correction factor, respectively. As expected, the absolute recalibration method provides better results than the relative one. In fact, the errors obtained with the absolute recalibration approach are almost equivalent to those obtained in the ideal case. The slight increase in the standard deviation comes from the deformation of the original database.

Table 2.4 summarizes the accuracy obtained in the form of mean estimation error plus one standard deviation. As can be seen, the loss of accuracy of the voltage and the temperature estimates after 10 years is only 14% and 25% with the proposed recalibration approach against a loss of 229% and 91% without recalibration. These results prove that the Multiprobe can be recalibrated against aging by only measuring the frequency shift of the seven ROs at a single known $\{V, T\}$ condition, even though it is quite a simple method. Additionally, this

Table 2.4: Summary of the estimation error of the VT estimation method under different aging situations.

	$\lvert\mu_{\epsilon_V}\rvert + \sigma_{\epsilon_V}$	$\lvert\mu_{\epsilon_T}\rvert + \sigma_{\epsilon_T}$
Fresh	$5.5mV$	$9.29\,^\circ C$
1 year	$12.9mV$ (+132%)	$14.54\,^\circ C$ (+56%)
10 years	$18.1mV$ (+229%)	$17.77\,^\circ C$ (+91%)
10y w/ recalibration	$6.3mV$ (+14%)	$11.66\,^\circ C$ (+25%)

correction can be performed periodically, especially in the first months, when the aging degradation is more important.

2.5 Conclusion

In this chapter, we have presented some of the numerous works done on adaptive techniques. From adaptation strategies to monitoring systems, the literature has a large range of solutions to on-line mitigate variability in digital circuits. However, the current adaptive techniques address mainly PVT variability, since aging effects are a relatively new issue. Moreover, as stated in Section 1.4, aging degradation strongly depends on the operating conditions. Thus, it is very difficult to accurately estimate it during the design phase.

We have also shown that existing aging sensors are mostly off-datapath monitoring techniques. They measure the degradation of a dedicated circuit, not the real degradation endured by the critical parts of the circuit. Invasive solutions such as slack time monitors, presented in Section 2.3.1, actually manage to detect the aging-induced delay shift of the critical paths. Nevertheless, they cannot distinguish the source of variation that originates the delay shift, which may also be due to voltage and temperature variations. Existing monitoring systems are not able to provide a reliable information about the circuit health. Note that such information is important, for example, to develop strategies aiming to increase the circuit lifetime.

In Section 2.3.2, we gave some examples of voltage and temperature monitors. For instance, [100] proposed a VT estimation method based on a small area sensor composed of only 7 ring-oscillators. However, no previous work had addressed the impact of aging on this kind of monitor so far. Therefore, in Section 2.4, we first demonstrated the accuracy loss induced by BTI and HCI effects on the VT estimation method of [100]. Then, we proposed an aging-aware recalibration approach. The developed approach consists basically in measuring the frequency shift (in Hertz) of each ring-oscillator in a known $\{V, T\}$ condition. This shift is then subtracted from all models in the database. For being a quite simplified method, this correction can be done regularly during the circuit lifetime making the sensor robust to aging variations.

In Chapter 3, we propose a new methodology to estimate at circuit-level the circuit f_{MAX} taking into account all sources of variation. The methodology results in two simplified models, one for the critical path delay and another one for the aging degradation. By feeding these models with voltage and temperature measurements, it is possible to on-line estimate the performance shift due not only to voltage and temperature variations, but also to aging. This can be done using the digital sensor from [100] and the recalibration method proposed in this chapter.

Chapter 3

Performance estimation under PVT and aging variations: a circuit-level methodology

Chapter 2 gave an overview of existing solutions to cope with variability in digital circuits. The different reviewed solutions range from sensors/monitors to adaptive strategies. These latter implement in-situ sensors that provide information about the chip state. Then, a *Sense & React* scheme is implemented in the circuit for dynamically adapting it against variations. However, the solutions presented do not distinguish between the different sources of variability, their goal being to adapt the circuit functioning point and reject variations that are seen as disturbances. Note that unlike temperature and voltage, aging induces irreversible variations that may lead the circuit to a permanent non-functional state. Moreover, the sensors are also prone to aging, leading to inaccurate information regarding the real state (Voltage, Temperature, Aging) of the chip.

As shown in Section 2.3.2, aging monitors are currently mainly based on ring-oscillators. However, they do not necessarily endure the same degradation as the functional parts of the circuit because aging degradation strongly depends on the circuit topology and on the workload which cannot be reproduced through a ring-oscillator.

This chapter proposes a new methodology for creating simplified but nonetheless accurate circuit-level models from existing device-level models in order to track aging in digital circuits. It consists of a model for the aging induced degradation and of another model for the path propagation delay (named hereafter *Delay*) encompassing the first one. Due to their low computational complexity, both models can be fed at runtime by voltage and temperature monitors to

provide the estimate of $Delay$ as well as its shift due to aging effects (named hereafter $\Delta Delay$). This methodology takes into account all factors that impact $Delay$ and aging, namely, the supply voltage V, the temperature T, the fabrication process, the circuit topology and the workload. The body bias Vbb could as well be taken into account without extra difficulty.

An advantage of the proposed methodology is that it is not invasive because there is no need to embed additional circuitry in the functional parts of the circuit. Moreover, instead of using canary structures such as ring oscillators, it provides information specifically about the circuit critical paths. As this methodology provides the resulting aging degradation for different operating conditions, it can also be used for other purposes. For instance, it can be used to estimate in advance the size of the safety margins necessary to ensure a safe operation of the circuit at the end of its lifetime, for all possible conditions of use.

This chapter is organized as follows. The objective of the proposed methodology is given in Section 3.1, while Section 3.2 summarizes the main steps of the proposed methodology. Sections 3.4 and 3.5 illustrate in details the application of the methodology on a benchmark circuit implemented in $28nm$ FD-SOI technology. This latter is described in Section 3.3. Finally, Section 3.6 validates the generated models through some use-cases and with another benchmark architecture.

The research activities presented in this chapter have been published in [7]. A patent application [8] was also filled in.

3.1 Objective

The objective of the proposed methodology is to abstract the complexity of existing aging models. So as to obtain low complexity circuit-level models that can be used either off-line or at run time to estimate the circuit degradation under different conditions of use (temperature, supply voltage, process, workload). The aging effects addressed here are the non-destructive ones, namely, BTI and HCI. Both phenomena change the transistors characteristics, in particular Vth, resulting in a larger propagation delay in digital circuits. In the last years, considerable efforts have been made to model BTI and HCI effects on transistors. Through experimental data measured on silicon, the aging-induced Vth shift is modelled with respect to variables such as the supply voltage, the temperature and the power-on time. The developed models can be then integrated in SPICE simulators that implement reliability functionalities, for instance Eldo, HSpice and SPECTRE.

Our proposal consists thus in a methodology to create a model for the circuit delay from existing device-level models as shown in Figure 3.1. The *Delay* is defined as the time between the rising edge clock and the data arrival at the input of the source flip-flop. The largest *Delay* in a circuit imposes its maximum functional frequency f_{MAX}. If a frequency higher than f_{MAX} is adopted, a setup time violation may occur leading to a faulty operation. The resulting *Delay* model incorporates not only aging variations, but also the PVT ones. This is particularly important for DVFS systems where the aging-induced f_{MAX} shift must be assessed for any V level.

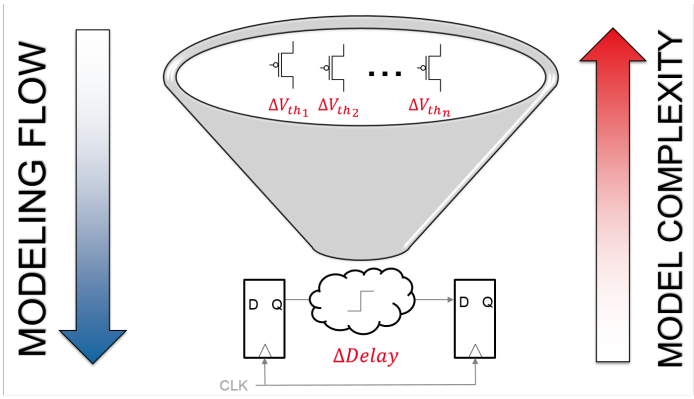


Figure 3.1: Aging model complexity abstraction from device-level to circuit-level.

3.2 Overview of the proposed methodology

The proposed methodology is illustrated in Figure 3.2. It is divided into two main stages, namely, *Delay Modelling* and *Aging Modelling*. The first stage produces a model to estimate the propagation delay of a circuit path under PVT variations. The second one generates a model that includes aging variations in the *Delay* model. Aging variations are represented as a shift in one of the parameters of the *Delay* model. Both models are constructed from the propagation delays simulated with SPICE. The following subsections detail both stages.

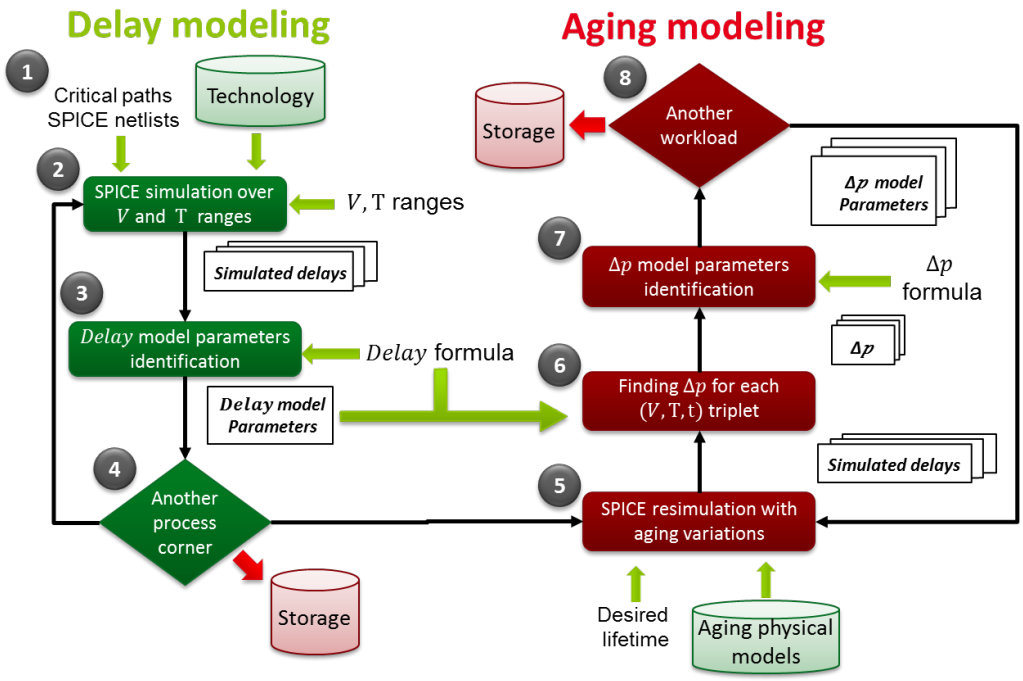


Figure 3.2: Two stages methodology proposed to obtain a circuit-level model of the path propagation delay, taking aging variations into account.

3.2.1 First stage: Delay Modelling

Each step of the *Delay Modelling* stage in Figure 3.2 is summarized below:

1. The netlists of one or more critical paths are extracted from the circuit design. Note that previous works address the selection of the aging-aware representative paths [117, 118]. Thus, the choice of how many paths and which ones must be selected is not addressed here.

2. The netlists are simulated within pre-defined V, T ranges using an electronic circuit simulator, e.g. SPICE [119]. The choice of V, T ranges depends on

the application specifications, while their respective steps $\Delta V, \Delta T$ depends on the desired modelling accuracy. The more V, T conditions are simulated, the more accurate the final models, but also the larger the simulation time. At the end, $n \times m$ propagation delays are gathered, where n and m are the number of V and T values, respectively.

3. The resulting propagation delays are fitted with a *Delay* formula that depends on V and T. The choice of the *Delay* formula is the most important step, since it will later be used for the *Aging Modelling* as well. The propagation delay could be simply translated into a simple polynomial expression. However, using an equation without any physical meaning would make the modelling of aging degradation much more intricate. Ideally, the *Delay* model should depends on Vth, which is known to be the transistor parameter shifted by BTI and HCI. The fitting process can be done through regression analysis using a numerical computing environment, e.g. MATLAB [120].

4. The previous steps are repeated for other process corners so that a set of parameters for the *Delay* equation is computed for each (path, corner) pair. The whole set of *Delay* model parameters are stored in a non-volatile memory. Some parameters may be only technology-dependent. Thus, they could be shared between all (path, corner) pairs. This would decrease the complexity of modelling several paths as well as the size of the memory used for model parameter storage.

3.2.2 Second stage: Aging Modelling

Each step of the *Aging Modelling* stage in Figure 3.2 is summarized below:

1. The critical path netlists are re-simulated with aging-induced variations. Several power-on times (t) are considered, as done for V and T. The aging physical models are either provided by the foundry or defined by the user. Note that models for BTI and HCI can be included together or only one of them can be considered. However, since an interplay exists between BTI and HCI effects [121, 122, 34], it is advisable to consider both of them at the same time. At the end of this step, $n \times m \times p$ propagation delays are generated, where n, m and p are the number of V, T and t values, respectively.

2. Aging degradation leads to an increase of the propagation delay. Therefore, there is at least one parameter of the *Delay* model that evolves with aging. This parameter shift Δp is evaluated by fitting the aged path delays with the *Delay* model and comparing the resulting parameters with the fresh ones. Vth being the transistor parameter most affected by aging [123], the

model parameter related to Vth should drift in a significant way. At the end of this step, a set of $n \times m \times p$ values of Δp are found, one for each (V, T, t) condition.

3. The $n \times m \times p$ values of Δp obtained are then fitted with an aging formula that depends on V, T and t. This formula can be based on classical device-level aging models (e.g. [123, 124, 35, 125, 126]), given that the dependences at a higher level should not be so different. The fitting process is similar to the one done in the *Delay Modelling* stage at step 3. The main difference is that there are now four dimensions instead of three, namely, Δp, V, T and t.

4. Finally, for a given application, the impact of the workload on aging is obtained using a cycle and bit accurate simulator tool, for instance Mentor Graphics Questa tool [127]). This tool provides both the circuit signal probabilities and the toggle rate for each circuit signal. The first one is a mandatory factor for estimating BTI degradation while the second one impacts mostly HCI. This procedure is repeated for as many workloads as needed. A set of Δp model parameters is obtained and stored for each workload. As for the process variations in the *Delay Modelling* stage, some parameters may be shared between several workloads. Thus decreasing the procedure complexity and the size of the memory used to store the model parameters.

3.3 Experimental set-up and SPICE reliability simulation

The capabilities of the proposed methodology are demonstrated in the rest of this chapter through an application example.

3.3.1 Benchmark circuit

The circuit under study has been implemented in STM $28nm$ FD-SOI technology. Here, a 32-bits Very Long Instruction Word (VLIW) DSP organized around a Multiplier-Accumulator (MAC) [61] is considered. This circuit is dedicated to Telecom applications. It is composed of a 10-stages pipeline, which allows reaching more than 1.5 GHz at the nominal supply voltage of 0.9V. Its high-level architecture is shown in Figure 3.3.

The SPICE netlists of its 300 most critical paths were extracted through Static Timing Analysis (STA) using Cadence's Encounter tool. One of them was randomly chosen to exemplify the proposed methodology. The chosen path is composed of 10 cells. It is depicted in Figure 3.4. The number after X stands for the cell drive strength. The abbreviations that appear on Figure 3.4 correspond to:

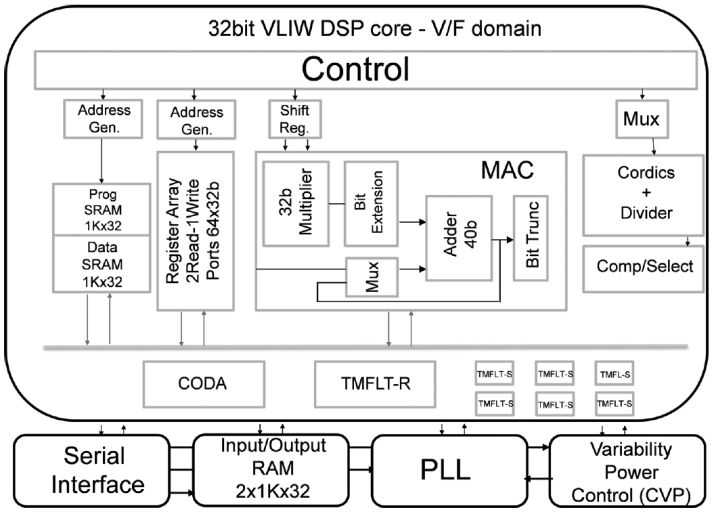


Figure 3.3: Architecture of the 32 bits VLIW DSP [61] used to validate the proposed methodology.

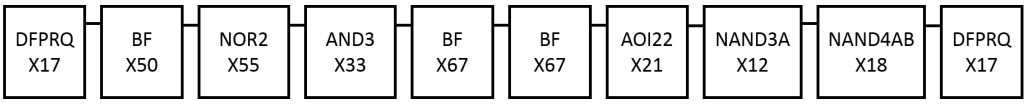


Figure 3.4: Composition of the critical path chosen here to demonstrate the capabilities of the proposed methodology.

- DFPRQ: Positive edge triggered non-scan D flip-flop with active low asynchronous reset;
- BF: Buffer;
- NOR2: 2 input NOR;
- AND3: 3 input AND;
- AOI22: Double 2 input AND into 2 input NOR;
- NAND3A: 3 input NAND with A input inverted;
- NAND4AB: 4 input NAND with A and B inputs inverted;

3.3.2 Device-level aging simulation

The Eldo User-Defined Reliability Model (UDRM) API was used to perform SPICE simulations with aging variations [116]. This API computes the stress experienced by each transistor during a transient simulation. Then, it runs a new simulation taking into account the resulting degradation. The reliability simulation flow is shown in Figure 3.5. During the first simulation, the circuit signals are set in order to reproduce the signal probabilities corresponding to a given workload. For the second simulation, they are set in order to make the signal

propagate from the source flip-flop to the destination one. As a consequence, we can assess the impact of aging on the propagation delay.

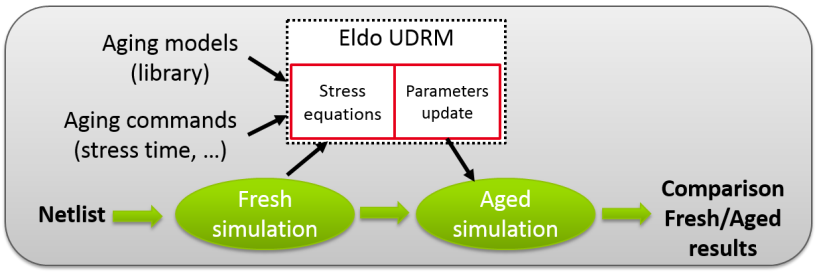


Figure 3.5: Diagram of the reliability simulation flow.

A state-of-the-art device-level model was adopted for both BTI and HCI effects [34]. This model takes into account an existing interplay between both phenomena. Note that traditional approaches consist in quantifying the impact of each mechanism separately. Then, the effects are considered additive. However, some defects created by both mechanisms in the gate oxide are actually the same. It follows that the actual combined degradation is smaller than adding both contributions independently. Figure 3.6 illustrates the pessimistic estimation performed by traditional approaches. As can be seen, the degradation estimated by the coupled model (▲) is close to the measurements (○) while the degradation estimated by adding both effects (●) is clearly pessimistic. Finally, the model adopted here also incorporates the BTI relaxation after stress.

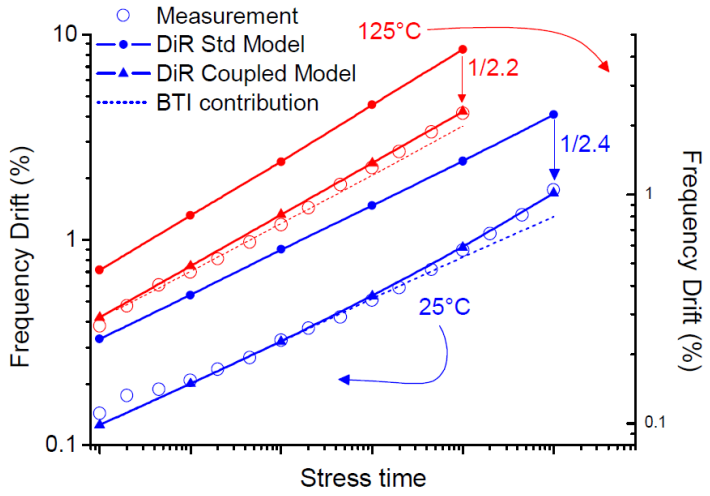


Figure 3.6: Ring oscillator frequency shift curves obtained with traditional BTI/HCI models (●, DiR Std Model) and with the coupled model (▲, DiR Coupled Model). Without the interplay between both phenomena, the models are much more pessimistic than the real degradation (○, Measurement) [34].

3.4 Application of the first stage of the proposed methodology: Delay modelling

Modelling the propagation delay is quite straightforward. All it requires is some SPICE simulations and data processing in a numerical computing environment. This can be fully automated through scripts.

The challenging point is the definition of the *Delay* formula itself. Basically, the propagation delay could be simply translated into a polynomial expression. However, it would be practically impossible to have only one equation parameter evolving with aging. It is more likely that all the equation parameters will shift from their "fresh" values, i.e. from their values when the circuit has not experienced aging. Moreover, both BTI and HCI effects impact the transistor threshold voltage Vth and carrier mobility. Thus, a *Delay* formula where both attributes are expressly represented must be considered. This can be achieved by relying on some theoretical assumptions.

3.4.1 Choice of the $Delay$ model formula

In 1990, T. Sakurai and R. Newton first proposed a simplified alpha-power law MOSFET model to estimate the propagation delay of an inverter [22, 128]. From this alpha-power law model, the propagation delay for a cell is expressed as:

$$Delay_{cell} \propto C_{out} * \frac{V}{I_d} \tag{3.1}$$

where C_{out} is the output load capacitance, V is the supply voltage and I_d is the drain current. The delay of a path is then simply the sum of the delays of all its cells. In this case, the supply voltage V is the same for all cells as well as the drain current I_d which is a technological parameter. Therefore:

$$\begin{aligned} Delay &\propto \sum C_{out} * \frac{V}{I_d} \\ &\propto C_{tot} * \frac{V}{I_d} \end{aligned} \tag{3.2}$$

where C_{tot} is the sum of all output load capacitances. As a consequence, the more cells in a path, the larger its propagation delay. The drain current is expressed as:

$$I_d \propto \mu(T) * (V - V_{th}(T))^{\alpha} \tag{3.3}$$

where $\mu(T)$ is the carrier mobility, $V_{th}(T)$ is the threshold voltage at temperature T and α is a positive constant related to the carrier velocity saturation. Finally, the propagation delay satisfies:

$$Delay(V,T) \propto C_{tot} * \frac{V}{\mu(T) * (V - V_{th}(T))^{\alpha}} \tag{3.4}$$

Both $\mu(T)$ and $V_{th}(T)$ decrease with increasing temperature. However, they impact the propagation delay in opposite ways, as shown in equation (3.4). This explains the Inverted Temperature Dependence (ITD) phenomenon. At low supply voltages, the threshold voltage dominates the drain current, while at high voltages it is determined by the carrier mobility. Therefore, the propagation delay increases with temperature increase at low supply voltages while it decreases at high supply voltages.

3.4.2 Estimation of the $Delay$ model formula parameters for the path under study

The path described in Section 3.3 is simulated over wide supply voltage and temperature ranges to generate a propagation delay surface. The surface is constructed by ranging V from 0.8V to 1.4V with a voltage step ΔV=20mV and T from $0°C$ to $150°C$ with step $\Delta T = 5°C$. A total of 961 points are simulated. Figure 3.7 shows the propagation delay surface as a function of the supply voltage V and the temperature T.

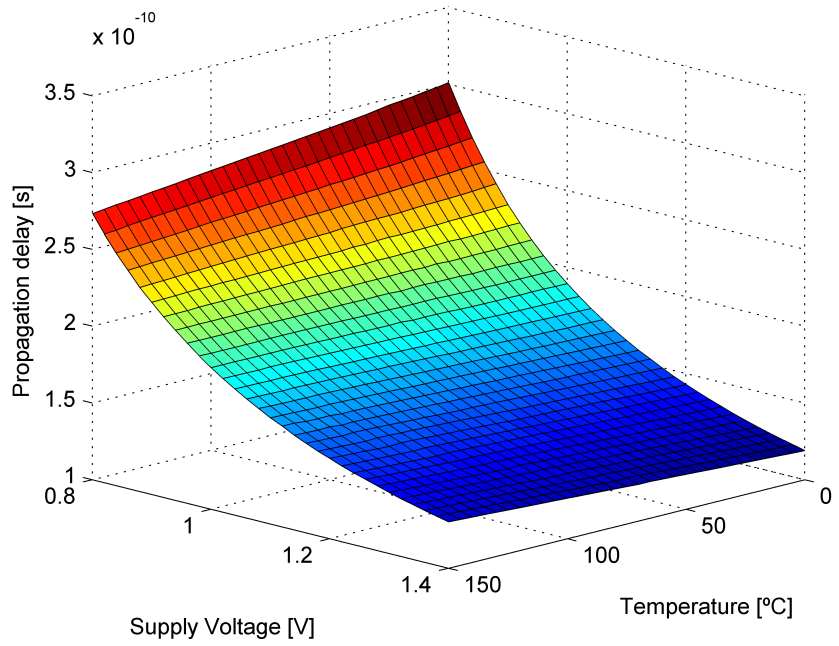


Figure 3.7: Propagation delay surface vs. supply voltage V and temperature T.

The delay surface is then separated in 31 iso-temperature $Delay(V)$ curves. Using equation (3.4), each curve is then fitted to the following equation:

$$Delay_T(V) = p_\beta + p_{\mu^{-1}}(T) * \frac{V}{(V - p_{V_{th}}(T))^{p_\alpha}} \tag{3.5}$$

where p_β is a constant while $p_{\mu^{-1}}$, $p_{V_{th}}$ and p_α correspond to $\frac{C_{tot}}{\mu(T)}$, $V_{th}(T)$ and α, respectively. The identification of the model parameters was performed with MATLAB using a non-linear least squares method (*nlinfit* command). For each value of T, a set of parameters is obtained.

Figure 3.8 shows the evolution of parameters p_β and p_α estimated for each temperature T. Both p_β and p_α were supposed to be temperature independent. From Figure 3.8, we can see that they can be considered constant (around $6.5 * 10^{-11}$ and 2.6, respectively) for T below $115°C$.

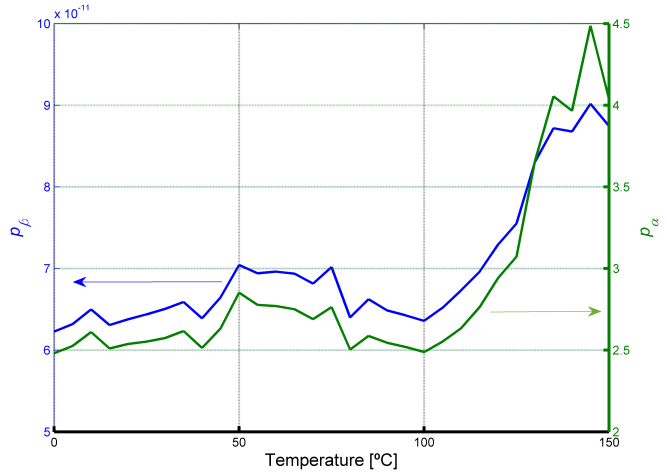


Figure 3.8: Evolution of parameters p_β (blue) and p_α (green) over temperature T.

Moreover, it must be noticed that they change together, which may mean that they have opposite effects on equation (3.5). To validate this assumption, the $Delay_T(V)$ curves are fitted again with equation (3.5), but with constant values for p_β and p_α. Figure 3.9 shows the mean normalized residuals for different combinations of values for different combinations of values for p_β and p_α. The presented residuals are calculated as follows:

$$\text{Residuals} = \underset{T}{\text{Mean}}(\underset{V}{\text{Mean}}(|\frac{\widehat{Delay_T}(V) - Delay_T(V)}{Delay_T(V)}|)) \tag{3.6}$$

where $Delay_T(V)$ is the path propagation delay obtained from SPICE simulation, $\widehat{Delay_T}(V)$ is the estimated value of delay in equation (3.5), $\underset{V}{\text{Mean}}$ is the mean value for all values of V and $\underset{T}{\text{Mean}}$ is the mean for all iso-temperature curves.

As it can be seen from Figure 3.9, the previous assumption is true. An increase of p_β is compensated by an increase of p_α without increase of the residual. p_β and p_α are finally fixed to $6.46 * 10^{-11}$ and 2.54, respectively. These values correspond to the smallest residual. Figure 3.10 depicts the evolution of $p_{\mu^{-1}}$ and $p_{V_{th}}$ over T after p_β and p_α have been fixed.

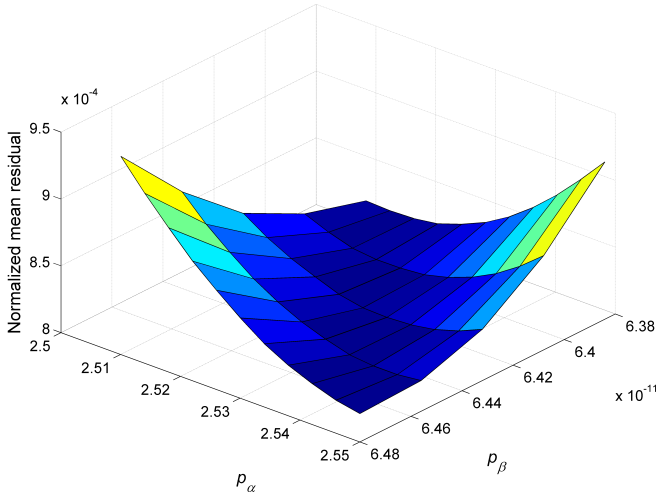


Figure 3.9: Normalized mean residuals (see equation(3.6)) obtained when fitting equation (3.5) with different values of p_β and p_α.

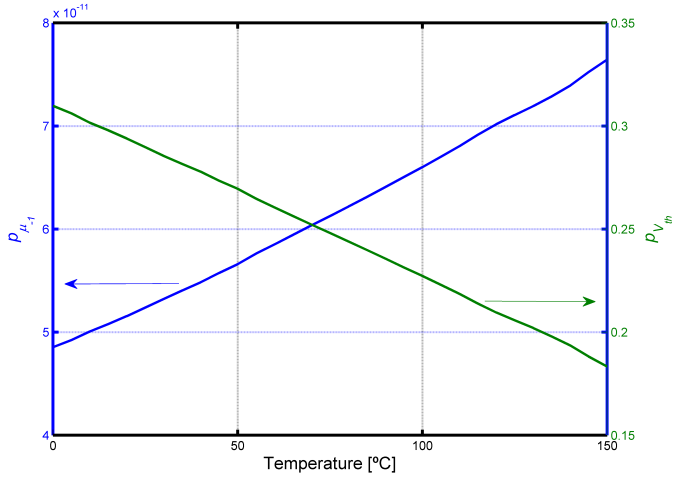


Figure 3.10: Evolution of parameters $p_{\mu^{-1}}$ (blue) and $p_{V_{th}}$ (green) vs. temperature T.

It can be seen that both $p_{\mu^{-1}}(T)$ and $p_{V_{th}}(T)$ curves possess similar behaviors but different slopes. They can be modelled as follows:

$$p_{\mu^{-1}}(T) = C_1 + k_1 T^{n_1}, k_1 > 0 \tag{3.7}$$

$$p_{V_{th}}(T) = C_2 - k_2 T^{n_2}, k_2 > 0 \tag{3.8}$$

By updating the equation (3.5) with the equation (3.7) for $p_{\mu^{-1}}(T)$ and the equation (eq:p2) for $p_{V_{th}}(T)$, we finally have a complete formula for $Delay(V, T)$.

The final formula has 8 parameters to be identified, as shown below:

$$Delay(V,T) = p_\beta + (C_1 + k_1 T^{n_1}) \frac{V}{(V - (C_2 - k_2 T^{n_2}))^{p_\alpha}} \tag{3.9}$$

where V is the supply voltage in Volt and T is the temperature in Kelvin.

The whole delay surface depicted in Figure 3.7 is fitted to equation (3.9) using MATLAB. Table 3.1 presents the parameters obtained and their respective 95% confidence interval CI. As can be seen, the confidence interval for each parameter is very small (¡1.3%) which means that the parameters are identified with a high degree of confidence. Also, this means that the model is not overparameterized. Indeed, when a model is over parameterized, IC is really large (¿10%) for some of its parameters.

Moreover, the fit between the estimated delay $Delay(V,T)$ and the measurements is really good, as can be seen in Figure 3.11 that shows the residual for each V,T point. The maximum residual found is 0.52% (0.68ps) at $(V,T) = (1.4V, 150°C)$ while the mean normalized error is 0.082%. The coefficient of determination R^2 is 0.99998. Note that R^2 determines how close the data are to the fitted regression surface. R^2 equals to 1 indicates that the regression surface perfectly fits the data. All these results validate the choice of equation (3.9) as $Delay$ model.

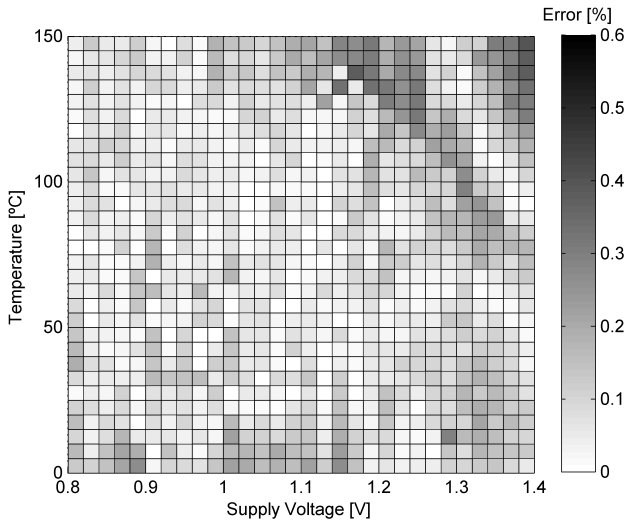


Figure 3.11: Map of normalized residual for the $Delay$ model in equation (3.9) versus temperature T and supply voltage V.

3.4.3 Analysis of the process variation on the delay estimation

The previous results have been obtained using *typical-typical (TT)* process corner during the simulations. This means that average values were adopted for the

Table 3.1: Parameters identified for the *Delay* model (see equation (3.9)) and their 95% confidence interval (CI).

Parameter name	Parameter value	Confidence interval
p_β	6.76e-11	0.07%
C_1	2.66e-11	0.24%
k_1	4.18e-16	1.24%
n_1	1.94	0.10%
C_2	0.46	0.13%
k_2	1.15e-4	1.24%
n_2	1.30	0.14%
p_α	2.68	0.03%

parameters of both NMOS and PMOS transistors. Besides *typical (T)* configuration, it is possible to choose between *slow (S)* and *fast (F)* ones, that represent the worst- and best-cases for global variations, respectively.

The delay surface depicted in Figure 3.7 is recomputed using *SS* and *FF* corners in order to include process variations in our proposed methodology. The new surfaces are used to estimate new sets of parameters for the *Delay* model given in equation (3.9). These new parameters are given in Table 3.2. Note that the confidence intervals are not reported because they are very small, as for the *TT* case. As can be seen, each *Delay* model parameter change from a process corner to another.

Table 3.2: Parameters of equation (3.9) for process corners *SS*, *TT* and *FF*, respectively.

	p_β	C_1	k_1	n_1	C_2	k_2	n_2	p_α
SS	6.96e-11	2.72e-11	6.60e-16	1.88	0.57	1.12e-4	1.30	2.60
TT	6.76e-11	2.66e-11	4.18e-16	1.94	0.46	1.15e-4	1.30	2.68
FF	6.88e-11	2.80e-11	7.98e-17	1.94	0.46	1.15e-4	1.30	2.68

However, a circuit after fabrication will not be exactly in one of these 3 process corners. It will be somewhere between the worst- (*SS*) and the typical-case (*TT*) or between the typical- and the best-case (*FF*). Therefore, a relationship between the *Delay* model parameters and the process corner must be constructed so that intermediary values can be found through interpolation. This is not possible with the parameters presented in Table 3.2 since the process corner dependency of the parameters is not straightforward. The reason may be that many parameters (in fact 8) appear in the model leading to an intricate dependency.

Yet, a dependency can be established if a reduced set of parameters is used. This can be accomplished by fixing the value of the other parameters for all corners. We must then find the parameters that are process independent and that

can be fixed at the mean value without loss of accuracy. For such, the relative standard deviation (RSD) of each parameter between the 3 process corners has been analyzed. RSD is simply the standard deviation divided by the mean. The parameter with the smallest RSD is fixed at its mean value. This procedure is repeated until the obtained residual get considerably higher than previous residuals. Table3.3 gives the fitting results for each iteration of this procedure. The mean normalized residual and the maximum one are reported.

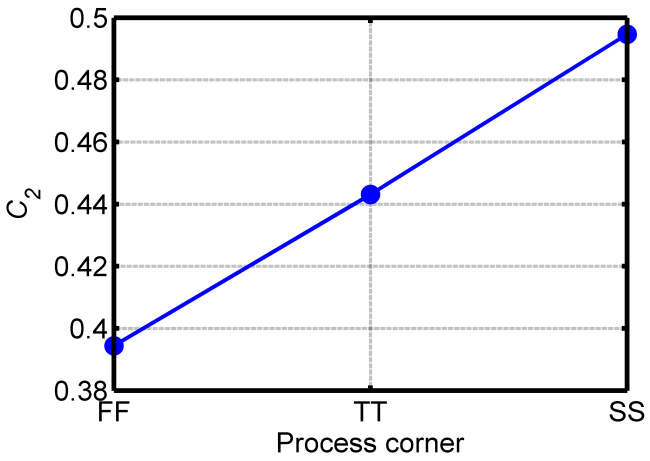
Table 3.3: Results of each iteration of the procedure used to fix some of the *Delay* parameters in equation (3.9) to a unique value for all process corners. Reported mean (Avg R) and maximum residual (Max R) are the average values for the 3 corners.

Iteration	1	2	3	4	5	6	7	8
Avg R (%)	0.10	0.10	0.10	0.10	0.11	0.12	0.13	1.38
Max R (%)	0.52	0.52	0.52	0.53	0.58	0.76	0.76	2.86
Coef fixed	p_β	C_1	n_2	k_2	n_1	k_1	p_α	-
Value	6.9e-11	2.8e-11	1.34	9.4e-5	2	3.1e-16	2.73	-
RSD (%)	2.23	2.37	2.65	4.94	4.79	0.98	7.93	-

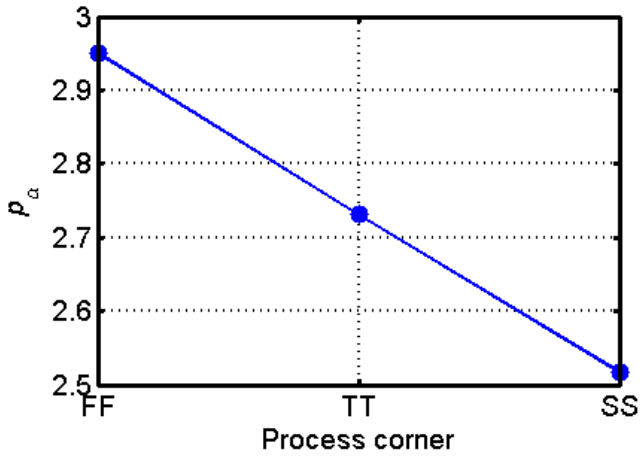
Until the 7^{th} iteration the residuals remained small even after fixing 6 parameters for all corners. It was only when p_α was fixed at the 8^{th} iteration that they significantly increased. Therefore, we can use the same parameter values for the 3 corners except for C_2 and p_α. Figure 3.12 shows their values for each corner. As can be seen, a linear evolution of C_2 (resp. p_α) versus the process corner can be considered. C_2 increases when the process corner becomes slower as expected since it is related to the threshold voltage, see equation(3.8), which is the transistor characteristic most affected by process variations. In turn, p_α decreases with a worse process corner resulting in a smaller drain current, see equation (3.3), and consequently, in slower transistors. Finally, from Figure 3.12, one can easily define the values of C_2 and p_α after circuit fabrication by using a linear interpolation along with a process calibration approach, some examples were previously given in Section 2.3.2.

3.5 Application of the second stage of the proposed methodology: Aging modelling

The *Aging Modelling* stage is quite similar to the *Delay Modelling* one, see Figure 3.2. It consists basically in SPICE simulations and data fitting after establishing the model. However, there is one additional step. As the propagation delay increases with the power-on time, there is at least one parameter of the *Delay* model that evolves with aging. We already know that both BTI and HCI phenomena impact mostly the threshold voltage but also the carrier mobility.



(a) C_2



(b) p_α

Figure 3.12: *Delay* model parameters C_2 and p_α in equation(3.9) for worst- (SS), typical- (TT) and best-case (FF) process corners.

Therefore, the first step of the *Aging Modelling* stage is to identify the parameter(s) of the *Delay* model that shifts with aging.

3.5.1 Shift of *Delay* model parameter(s) due to aging

Another delay surface has been generated to determine the parameter(s) of the *Delay* model that shifts with aging. This new surface has been created with SPICE simulations taking into account aging variations, as previously described

in Section 3.3.2. The conditions of stress adopted for this case were 1.2V, 125°C and 20 years. A first simulation was performed at 1.2V and 125°C to produce the stress stimuli which are used by the simulator to compute the resulting degradation of each transistor extrapolated to a power-on-time (t) of 20 years. A library was generated containing all the parameters of the degraded transistors. This library was then used to construct an aged delay surface, allowing the same procedure as for the "fresh" condition, with V ranging from 0.8V to 1.4V with step $\Delta V = 20mV$, and with T evolving from 0°C to 150°C with step $\Delta T = 5°C$.

The *Delay* model of equation (3.5) was identified with the new set of simulated path delays to observe how the parameters shift. The parameters p_β and p_α were fixed to their fresh values (see Table 3.1) because we are interested in the ones related to the carrier mobility and the threshold voltage. Figure 3.13 compares the fresh and aged values for both $p_{\mu^{-1}}(T)$ and $p_{V_{th}}(T)$. As we can see, there is a little decrease of the mobility ($p_{\mu^{-1}} \propto 1/\mu$) while a significant increase of the threshold voltage is observed ($p_{V_{th}}(T) \propto V_{th}$). The proposed methodology could be applied with two parameters that shift over time, but it would become considerably more complex. First, an extra model would be needed for the second parameter shift. Next, more than one simulated aged delay would be necessary for each (V, T, t) condition to find both shifted parameters. Therefore, for simplification purposes, only the $p_{V_{th}}$ shift is considered in the model. This parameter shift will be called $\Delta p_{V_{th}}$. Thus, the final *Delay* model including aging variations is expressed as:

$$Delay(V,T,t) = p_\beta + p_{\mu^{-1}}(T)\frac{V}{(V - (p_{V_{th}}(T) + \Delta p_{V_{th}}(V,T,t)))^{p_\alpha}} \tag{3.10}$$

where $\Delta p_{V_{th}}(V,T,t)$ represents both BTI and HCI effects coupled.

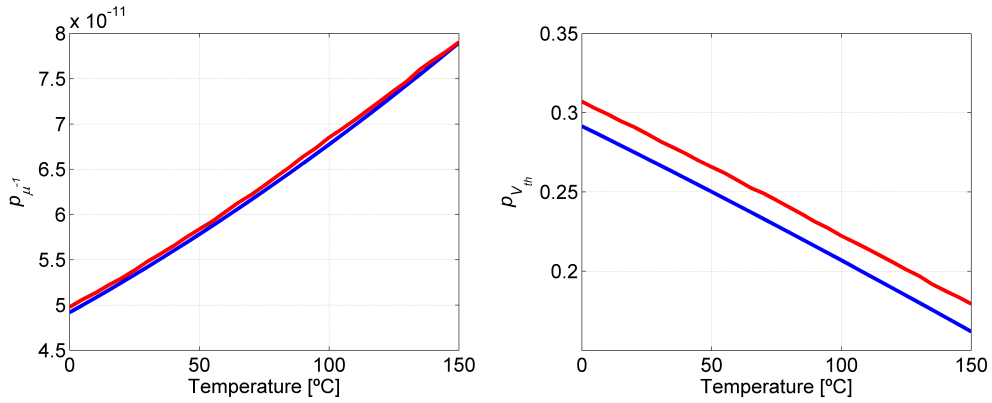


Figure 3.13: Original (blue) and aged (red) values of $p_{\mu^{-1}}(T)$ (left) and $p_{V_{th}}(T)$ (right).

The next step consists in calculating the $\Delta p_{V_{th}}$ for various V, T and t conditions in order to create a $\Delta p_{V_{th}}(V,T,t)$ model. Stress stimuli were generated

for $31 \times 31 \times 30$ different (V, T, t) conditions of use (28830 in total). Then, they were used to simulate the aged path delays. The ranges adopted were $[0.8, 1.4]V$, $[0, 150]°C$ and $[0, 20]$ years. The $\Delta p_{V_{th}}$ for each (V, T, t) condition was estimated using the MATLAB *fzero* command. This command computes $\Delta p_{V_{th}}$ where $fun(\Delta p_{V_{th}}) = 0$. $fun(\Delta p_{V_{th}})$ is the difference between the simulated aged delay and equation (3.10), as follows:

$$f(\Delta p_{V_{th}}) = |Delay(V_i, T_i, t_i)_{SPICE} - Delay(V_i, T_i, \Delta p_{V_{th}})_{Model}| \tag{3.11}$$

Figure 3.14 shows three examples of $\Delta p_{V_{th}}(V, T, t)$. Actually $\Delta p_{V_{th}}(V, T, t)$ being a 4-dimensions surface, one of the 3 variables is fixed to generate the 3-dimensions surfaces. The blank spaces correspond to situations when the computed value of $\Delta p_{V_{th}}$ was negative, i.e. the aged path delay smaller than the fresh one. This situation appears because a SPICE transient simulation is indeed the numerical resolution of an algebraic differential system of equations. As such, it is not an exact procedure and it is therefore susceptible to some accuracy errors. In total, 236 out of 28830 (V, T, t) conditions of use (0.82%) were ignored because they resulted in a negative numerical value for $\Delta p_{V_{th}}$.

3.5.2 Construction of the $\Delta p_{V_{th}}$ model formula

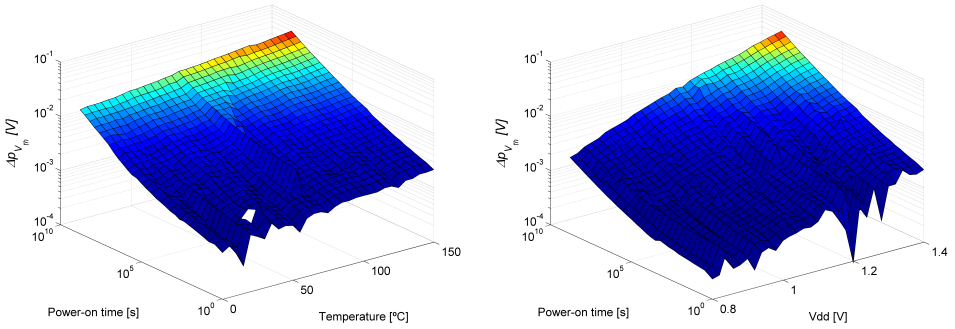
The next step in the *Aging Modelling* stage is to fit the $\Delta p_{V_{th}}(V, T, t)$ computed values to a model. As in the *Delay Modelling* stage, the definition of the $\Delta p_{V_{th}}$ model is the most important step in the *Aging Modelling* stage. It is even more complex than the *Delay* model since it has an extra dimension, namely, the power-on time.

As for the *Delay*, the $\Delta p_{V_{th}}$ model can be built through an empirical method based on some theoretical assumptions. Existing BTI models from the literature [123, 124, 35, 125, 126] are used as template for modelling the temperature, voltage and time dependencies of $\Delta p_{V_{th}}$. BTI models are preferred over HCI ones because the aging degradation in digital circuits is mostly due to BTI [29, 129]. Nevertheless, as $\Delta p_{V_{th}}$ is the resulting degradation at circuit-level of both effects coupled, the final model will be defined so that the best fitting results are attained, i.e. the smallest residuals and the narrowest confidence intervals possible are obtained.

The definition at first of the complete expression of the $\Delta p_{V_{th}}(V, T, t)$ model is in practice very difficult. Instead, it is possible to evaluate the contribution of each variable one at a time.

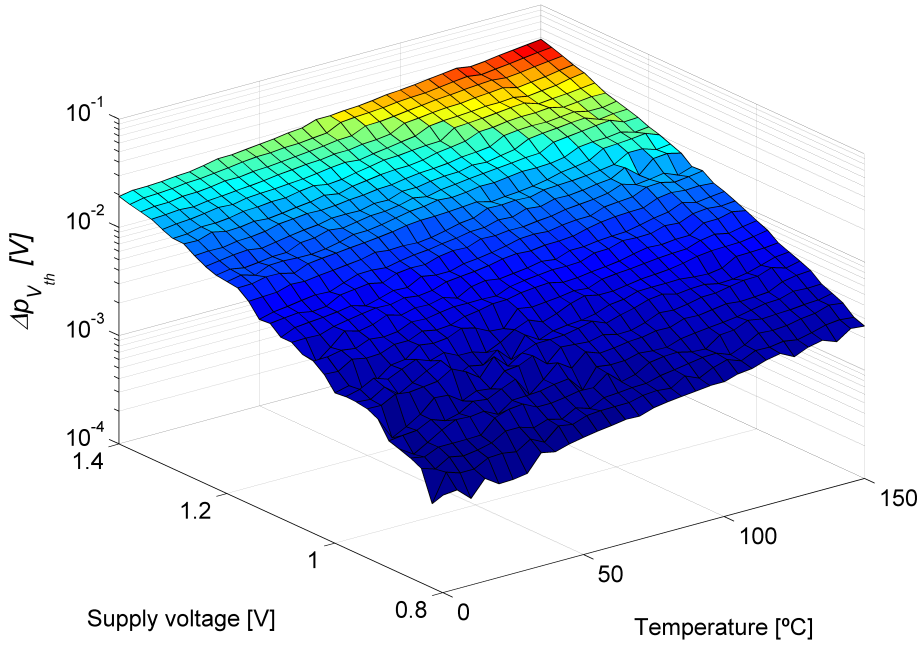
Temperature dependence of $\Delta p_{V_{th}}$

Among them, the variable which dependency is widely accepted in the literature is the temperature. It is often assumed that the BTI dependence on temperature



(a) $\Delta p_{V_{th}}(V,T,t)$ with $V = 1.4V$

(b) $\Delta p_{V_{th}}(V,T,t)$ with $T = 150°C$



(c) $\Delta p_{V_{th}}(V,T,t)$ with $t = 20$ years

Figure 3.14: $\Delta p_{V_{th}}$ surfaces extracted from the aged path delays.

follows an Arrhenius law [123]:

$$\Delta p_{V_{th}}(T) \propto e^{-E_a/kT} \tag{3.12}$$

where E_a is the temperature activation energy in Joules, k is the Boltzmann's constant ($8.617e5eV/K$) and T is the temperature in Kelvins. To estimate the

value of E_a, the $\Delta p_{V_{th}}$ data set has been cut in $31 \times 30 \Delta p_{V_{th}}(T)$ curves, one for each (V, t) couple. A logarithmic function has then be applied to each curve. Therefore, the fitting can be performed as follows:

$$log(\Delta p_{V_{th}}(T)) = A - \frac{E_a}{kT} \tag{3.13}$$

where A is a coefficient depending on V and t.

Figure 3.15 shows E_a obtained for each $\Delta p_{V_{th}}(T)$ curve. We can see that E_a stays around an average value of 0.0775 ($E_a/k = 900$) in most cases.

Figure 3.16 compares the evolution of $\Delta p_{V_{th}}$ over T for $(V, t) = (1.4V, 20$ years) obtained through simulation and built with equation (3.12). We can see that the simulated data follows quite well the Arrhenius law.

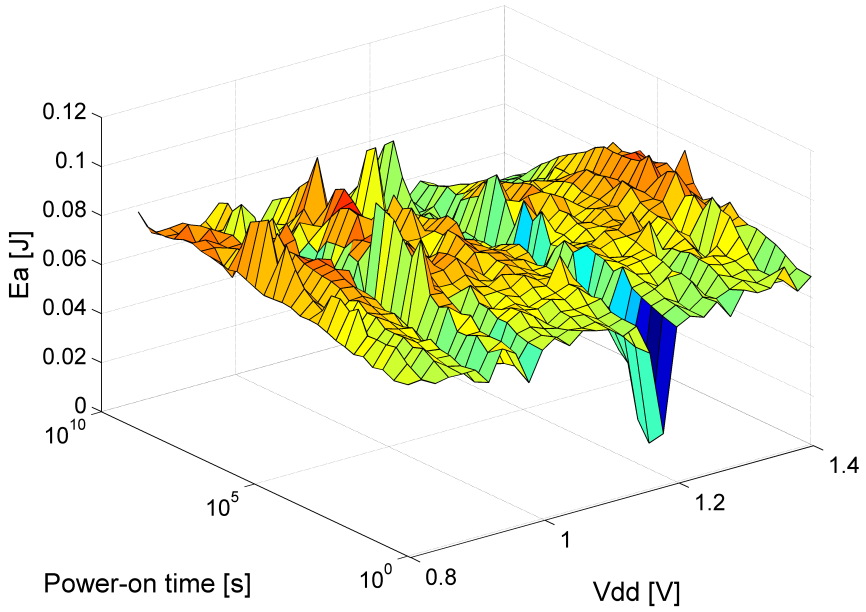


Figure 3.15: Values of E_a found by fitting $\Delta p_{V_{th}}(T)$ curves into equation (3.13). The average value of E_a is 0.0775.

Supply voltage dependence of $\Delta p_{V_{th}}$

The contribution of temperature being defined, it is now possible to determine the supply voltage contribution through $\Delta p_{V_{th}}(V, T)$ surfaces. The BTI dependence on the supply voltage is usually expressed through an exponential function [125, 130, 124]:

$$\Delta p_{V_{th}}(V, T) = A * V^{\gamma} * e^{-900/T} \tag{3.14}$$

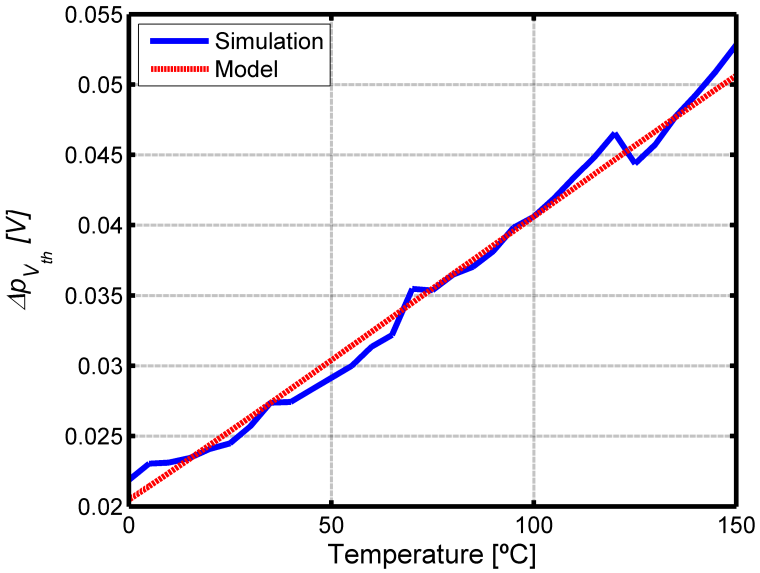


Figure 3.16: $\Delta p_{V_{th}}(T)$ for $(V, t) = (1.4V, 20$ years). Blue: Simulation. Red: Model (Arrhenius law).

In other cases it is expressed through a power function (V^{γ}) [31, 131]:

$$\Delta p_{V_{th}}(V, T) = A * e^{\gamma V} * e^{-900/T} \tag{3.15}$$

The $\Delta p_{V_{th}}$ data set was cut in 30 (V, T) surfaces, one for each value of t. These surfaces were fitted using both equations (3.14) and (3.15), where A is a coefficient depending on t. The average Root Mean Square Error (RMSE) and the average 95% confidence intervals of the parameters are reported in 3.4. Note that the RMSE is the square root of the variance of the residuals. It is an absolute measure of fit (same unit)and a lower RMSE value indicates a "better" fit. In turn, the Confidence Intervals are used to assess the quality of the estimate of the parameters. The narrower the confidence interval, the more precise the estimate is. From Table 3.4, it is possible to see that both equations presented quite similar fitting results. Therefore, no conclusion about the supply voltage dependence of $\Delta p_{V_{th}}$ model could be drawn from these results. Both equations 3.14) and (3.15) are kept for the following analysis (power-on time).

Table 3.4: Average fitting results of the 30 $\Delta p_{V_{th}}(V, T)$ surfaces using the equations (3.14) and (3.15).

Equation	γ	RMSE	95% Confidence Interval
Power (V^{γ})	4.95	5.14e-4	±2.40%
Exponential ($e^{\gamma V}$)	4.13	5.38e-4	±2.32%

Power-on time dependence of $\Delta p_{V_{th}}$

Finally, the time-dependence was modelled after both temperature and voltage contributions were analyzed. This was achieved by using the whole data set. For the BTI time-dependence, two theories are usually accepted, namely, Reaction-Diffusion [132], where it is expressed as a power function (t^n), and Charge-Trapping [133], where a logarithmic function is adopted instead ($log(1+nt)$). Both of them were inserted in the equations (3.14) and (3.15). Therefore, the $\Delta p_{V_{th}}$ data were then fitted to the 4 following models:

$$\Delta p_{V_{th}}(V,T,t) = C * t^n * V^\gamma * e^{-900/T} \tag{3.16}$$

$$\Delta p_{V_{th}}(V,T,t) = C * t^n * e^{\gamma V} * e^{-900/T} \tag{3.17}$$

$$\Delta p_{V_{th}}(V,T,t) = C * log(1+nt) * V^\gamma * e^{-900/T} \tag{3.18}$$

$$\Delta p_{V_{th}}(V,T,t) = C * log(1+nt) * e^{\gamma V} * e^{-900/T} \tag{3.19}$$

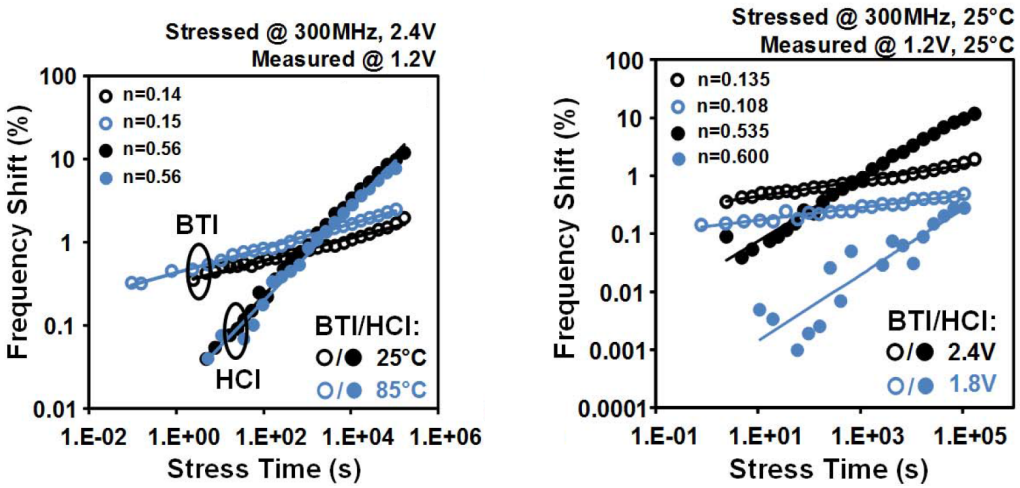
Table 3.5: Identification fitting results for $\Delta p_{V_{th}}$ data to the equations (3.16 - 3.19). RMSE is the Root Mean Square Error while Max R stands for the maximum value of the residual.

Equation	Eq. (3.16)	Eq. (3.17)	Eq. (3.18)	Eq. (3.19)
RMSE	3.12e-4	3.47e-4	5.81e-4	5.90e-4
Max R	9.4e-3	8.0e-3	30.7e-3	29.6e-3
γ (95% CI)	4.88 (0.2%)	4.27 (0.2%)	4.05 (0.5%)	3.59 (0.4%)
C (95% CI)	2.5e-3 (0.5%)	3.4e-5 (1.0%)	2.0e-3 (0.6%)	5.6e-5 (2.1%)
n (95% CI)	0.16 (0.1%)	0.16 (0.2%)	0.23 (4.9%)	0.21 (4.9%)

Table 3.5 gives the fitting results for the 4 models. The RMSE, the maximum residual and the 95% Confidence Interval CI of each parameter are also reported besides the parameters themselves. By comparing the first two models with the last two ones, we conclude that the $\Delta p_{V_{th}}$ data fits better to a t^n time-dependence. Nevertheless, the voltage contribution still remains hard to model since the two proposed expressions conduct to similar results. The authors in [35] modelled the contributions of both BTI and HCI effects on the frequency shift of ring-oscillators through experimental data. Moreover, they extracted the time component of both phenomena under different temperatures and stress voltages, as shown in Figure 3.17. From these graphs it is possible to note that BTI and HCI have completely different time dynamics. Moreover, while the temperature do not alter the BTI/HCI time components, the supply voltage does.

Based on the findings in [35], we decided to evaluate a power law with two time exponents for the time dependence. At first, we tested it with two constant exponents:

$$\Delta p_{V_{th}}(V,T,t) = (C_1 * t^{n_1} + C_2 * t^{n_2}) * V^\gamma * e^{-900/T} \tag{3.20}$$



(a) (b)

Figure 3.17: BTI and HCI contribution on induced frequency shift under different (a) temperatures and (b) supply voltages [35].

$$\Delta p_{V_{th}}(V,T,t) = (C_1 * t^{n_1} + C_2 * t^{n_2}) * e^{\gamma V} * e^{-900/T} \tag{3.21}$$

Then, we adopted time exponents with a logarithmic dependence on the supply voltage:

$$\Delta p_{V_{th}}(V,T,t) = (C_1 * t^{n_1+a_1*log(V)} + C_2 * t^{n_2+a_2*log(V)}) * V^{\gamma} * e^{-900/T} \tag{3.22}$$

$$\Delta p_{V_{th}}(V,T,t) = (C_1 * t^{n_1+a_1*log(V)} + C_2 * t^{n_2+a_2*log(V)}) * e^{\gamma V} * e^{-900/T} \tag{3.23}$$

Table 3.6 shows the RMSE, the maximum residual and the obtained parameters with their respective 95% Confidence Intervals CI for the parameter identification performed for models in equations (3.20 - 3.23). Comparing the results in the first two columns with the ones in Table 3.5 it is possible to see that equations with 2 time exponents produced better fitting results. The RMSE has been reduced by around 25% while the maximum residual became more than 40% smaller than before. The wider Confidence Intervals are acceptable since there are more parameters in the equation.

The same conclusion stands when we analyze the fitting results for equations with voltage-dependent time exponent in the last two columns. The RMSE has been reduced by around 20% while the confidence intervals increase when compared to the first two columns. Besides, there is a slight improvement when using a power function (equations (3.20) and (3.22)) for the voltage dependence instead of an exponential function (equations (3.21) and (3.23)).

Table 3.6: Fitting results of $\Delta p_{V_{th}}$ using models with two time components. Constant exponents were used in equations (3.20) and (3.21), while voltage-dependent ones were adopted in equations (3.22) and (3.23).

Equation	Eq. (3.20)	Eq. (3.21)	Eq. (3.22)	Eq. (3.23)
RMSE	2.28e-4	2.67e-4	1.84e-4	2.08e-4
Max R	5.0e-3	4.7e-3	4.7e-3	9.3e-3
γ (95% CI)	5.05 (0.1%)	4.33 (0.1%)	3.93 (1.4%)	3.09 (1.2%)
C_1 (95% CI)	3.4e-3 (0.7%)	4.5e-5 (1.0%)	4.5e-3 (1.4%)	2.3e-4 (4.8%)
C_2 (95% CI)	3.5e-4 (6.2%)	3.5e-6 (7.8%)	4.0e-4 (4.9%)	1.1e-5 (3.7%)
n_1 (95% CI)	0.07 (3.7%)	0.08 (3.4%)	0.05 (5.3%)	0.06 (0.8%)
n_2 (95% CI)	0.25 (1.0%)	0.26 (1.2%)	0.24 (0.8%)	0.27 (5.3%)
a_1 (95% CI)	-	-	0.08 (9.3%)	0.11 (5.3%)
a_2 (95% CI)	-	-	0.06 (2.8%)	0.10 (1.5%)

Finally, we can conclude that equation (3.22) is the most accurate model for $\Delta p_{V_{th}}$. Basically, it gives the "best" fitting results and it is in accordance with traditional BTI/HCI models found in the literature. However, equation (3.20) could be adopted to have a less complex model despite some accuracy loss. There exist an interest of adopting the most accurate equation for the *Delay* model in order to find the parameter shift due to aging. In contrast, for the $\Delta p_{V_{th}}$ model a better trade off between accuracy and complexity may be sought instead. A less complex model can be a better choice in particular when an online implementation of the models is envisaged.

3.5.3 Example of $\Delta p_{V_{th}}$ model parameters

The $\Delta p_{V_{th}}$ data is then fitted using following equation:

$$\Delta p_{V_{th}}(V,T,t) = (C_1 * t^{n_1+a_1*log(V)} + C_2 * t^{n_2+a_2*log(V)}) * V^{\gamma} * e^{-E_a/kT} \quad (3.24)$$

This model contains 8 parameters to be estimated from 28830 values for $\Delta p_{V_{th}}$. The ranges adopted for V, T and t are $[0.8, 1.4]V$, $[0, 150]°C$ and $[0, 20]$ years, respectively, with $\Delta V = 2mV$, $\Delta T = 5°C$ and 30 logarithmically distributed values for t.

The identified parameters with their respective Confidence Intervals CI are given in Table 3.7. The RMSE of fitting was $0.18mV$ while the maximum residual was $4.8mV$. Figure 3.18 shows the cumulative distribution function (CDF) of the residuals. It is possible to see that almost 95% of the errors stayed within $\pm 1mV$.

Figure 3.19 compares three $\Delta p_{V_{th}}$ surfaces obtained through SPICE simulations to the surfaces constructed with the proposed model. The small fitting residuals together with the similarity between the surfaces validate the chosen $\Delta p_{V_{th}}$ model and the parameters that have been identified.

Table 3.7: $\Delta p_{V_{th}}$ model parameters for equation (3.24).

γ (95% CI)	3.93 (1.4%)
E_a/k (95% CI)	914.5 (0.1%)
C_1 (95% CI)	4.7e-3 (1.4%)
C_2 (95% CI)	4.3e-4 (5.0%)
n_1 (95% CI)	0.05 (5.5%)
n_2 (95% CI)	0.24 (0.8%)
a_1 (95% CI)	0.087 (8.7%)
a_2 (95% CI)	0.060 (2.8%)

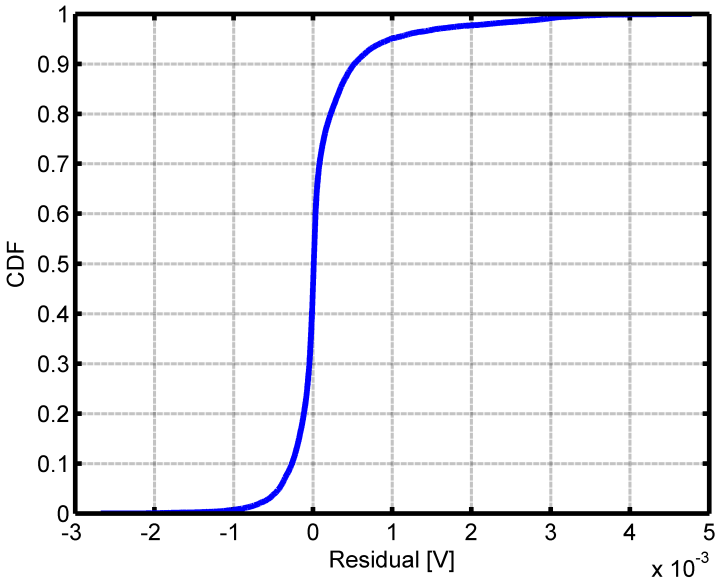
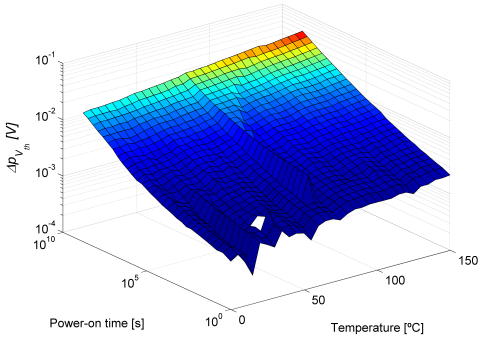
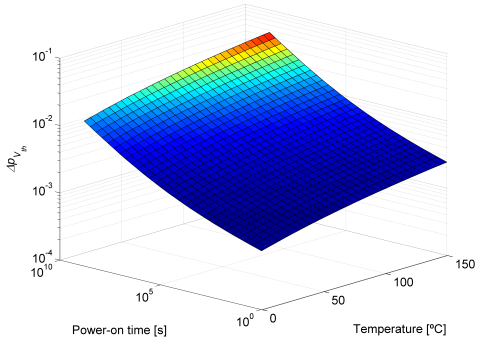


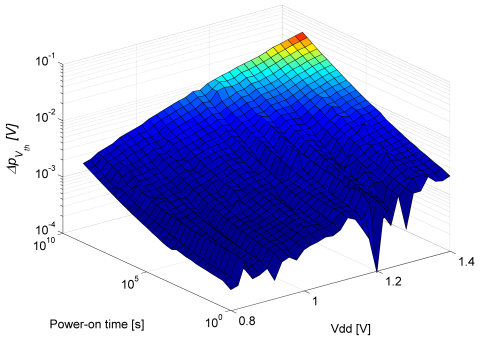
Figure 3.18: Cumulative distribution function of residuals after fitting $\Delta p_{V_{th}}$.



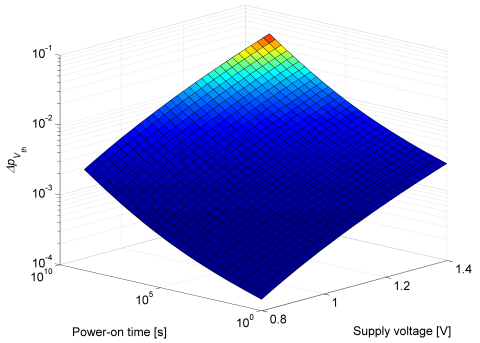
(a) $\Delta p_{V_{th}}(T,t)$ for $V = 1.4V$ built with simulation



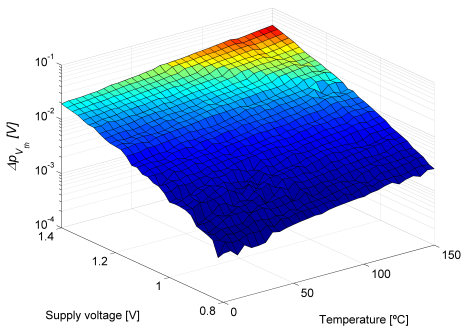
(b) $\Delta p_{V_{th}}(T,t)$ for $V = 1.4V$ built with the proposed model



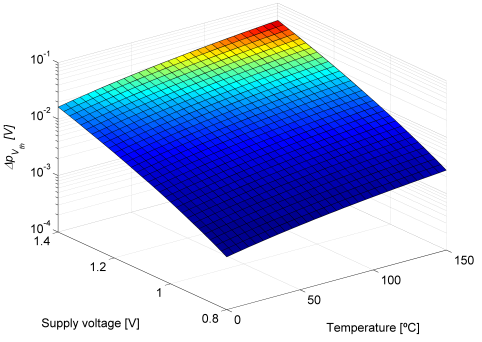
(c) $\Delta p_{V_{th}}(V,t)$ for $T = 150°C$ built with simulation



(d) $\Delta p_{V_{th}}(V,t)$ for $T = 150°C$ built with the proposed model



(e) $\Delta p_{V_{th}}(V,T)$ for $t = 20$ years built with simulation



(f) $\Delta p_{V_{th}}(V,T)$ for $t = 20$ years built with the proposed model

Figure 3.19: Comparison of $\Delta p_{V_{th}}$ surfaces generated through SPICE simulation and the proposed model.

3.5.4 Impact of workload on aging

The temperature and the supply voltage are not the only factors that influence the aging degradation of a circuit. The workload is as relevant as them for the final delay shift. By workload we mean the application running in the circuit along with the input data that define the activity and the signal probabilities of each net. For instance, a PMOS transistor whose gate voltage is almost all the time at V will not get stressed. On the other hand, it will endure the worst degradation possible if its gate is stuck at ground. In a circuit path, all cells are connected through a main signal which is propagate from the output of the first flip-flop to the input of the second flip-flop. Other signals are connected to the cells with 2 or more inputs. In the path used as example here (10 cells long) 12 additional signals exist. All these signals are as important as the main one for computing the aging-induced delay shift. An experiment has been performed through SPICE simulations where the delay shift was measured for different signal configurations. In each configuration, all the signal probabilities changed except for the main one. Three situations were tested:

- *Normal*, where the secondary signals were set in order to propagate the main one;
- *Inverted*, where the signals were inverted from their previous values;
- *50/50*, where they were half the time at V and the other half at ground.

Table 3.8 shows the path delay shifts for both fall and rise transitions with a stress condition of $1.2V$ and $125°C$ extrapolated to 20 years. Different delay shifts were observed depending on the signals configuration. Moreover, the aging degradation followed distinct trends according to the signal edge.

Table 3.8: Delay shift due to different configurations for the secondary signals. Stress conditions = $(1.2V, 125°C, 20$ years).

	Fresh	*Normal*	*Inverted*	*50/50*
Fall	$150.58ps$	$155.29ps$ (3.13%)	$154.22ps$ (2.42%)	$153.80ps$ (2.14%)
Rise	$140.17ps$	$145.33ps$ (3.68%)	$144.46ps$ (3.06%)	$145.28ps$ (3.65%)

It is impossible to include the workload as a variable in the $\Delta p_{V_{th}}$ model due to its numerous possibilities of signal combinations. The solution therefore is to re-estimate the $\Delta p_{V_{th}}$ model parameters for as many workloads as needed. A similar procedure to the one done for process variations in the *Delay* model has therefore to be applied. For each workload, the signal probabilities of each transistor is obtained through cycle accurate RTL simulations. Then the netlist is simulated with the corresponding signal probabilities, producing a new degradation. As for process variations on the *Delay* model, some of the $\Delta p_{V_{th}}$ parameters may be

shared between different workloads if their values does not significantly vary with the workload.

Three sets of random signal probabilities were generated to simulate different workloads. Simulations with aging variations were performed within V, T and t ranges of $[0.8, 1.4]V$, $[0, 150]°C$ and $[0, 20]$ years, respectively, with $\Delta V = 5mV$, $\Delta T = 15°C$ and 10 logarithmically distributed values for t. In total, 1430 (V, T, t) conditions of use were simulated ($10 \times 13 \times 11$). The 8 $\Delta p_{V_{th}}$ parameters in equation (3.24) were obtained for each workload as shown in Table 3.9. As expected, the value of each parameter changes from a workload to another.

Table 3.9: Sets of parameters in equation (3.24) computed for different signal probabilities.

	Workload A	Workload B	Workload C
γ	4.21	4.54	3.16
E_a/k	884.3	804.6	662.0
C_1	4.7e-3	2.7e-3	2.3e-3
C_2	4.1e-4	2.4e-4	3.1e-5
n_1	0.040	0.068	0.054
n_2	0.243	0.266	0.307
a_1	0.109	0.050	0.088
a_2	0.044	0.028	0.093

Figure 3.20 depicts the 3 curves of $\Delta p_{V_{th}}$ over time for a stress condition of $1.4V$ and $150°C$. As it can be seen, there is a considerable change in $\Delta p_{V_{th}}$ for Workload A and Workload C, the first one being more than 2 times greater than the second one. This shows the importance of taking the workload into account for the aging estimation.

Finally, some of the parameters can be shared between all workloads if a process similar to the one adopted in Section 3.4.3 for process variations on the *Delay* model is applied. It reduces the memory required for storing the parameters and eases the $\Delta p_{V_{th}}$ estimation since some parts of the equation do not need to be recomputed between different workloads. The relative standard deviation (RSD) of each parameter between the 3 sets is analyzed. The one with the smallest RSD is then fixed at its mean value. This procedure is repeated until the mean RMSE and the mean maximum residual considerably increase. Table 3.10 reports the results for each iteration.

It is possible to see that until the sixth iteration there was almost no loss of accuracy after fixing 5 parameters of equation (3.24). It was only when C_2 was fixed at iteration 7 that both the RMSE and the maximum residual finally became significantly deteriorated. Therefore, in this example, 5 parameters (E_a/k, γ, C_1, n_2 and a_2) can be shared between all workloads while only 3 parameters (n_1, a_1 and C_2) have specific values for each workload.

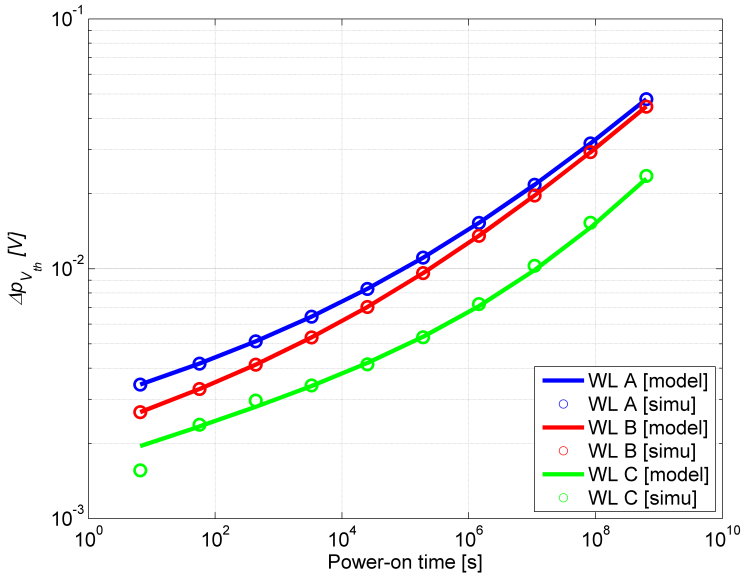


Figure 3.20: $\Delta p_{V_{th}}$ evolution over time for 3 workloads (stress condition of $1.4V$ and $150°C$). The $\Delta p_{V_{th}}$ obtained in simulation are represented by circles while the lines correspond to the model with parameters shown in Table 3.9.

Table 3.10: Results of each iteration to find the $\Delta p_{V_{th}}$ parameters (see equation (3.24)) that can be shared between all workloads. Reported RMSE and maximum residual (Max R) are the average values.

Iteration	1	2	3	4	5	6	7
RMSE	4.63e-4	4.86e-4	4.87e-4	4.88e-4	4.91e-4	4.91e-4	5.58e-4
Max R	3.74e-3	3.78e-3	3.78e-3	3.77e-3	3.75e-3	3.74e-3	4.63e-3
Coef fixed	E_a/k	γ	n_2	C_1	a_2	C_2	-
Value	815.6	4.0	0.267	3.5e-3	0.05	1.6e-4	-
RSD	14.4%	17.0%	14.9%	14.7%	17.6%	20.3%	-

3.5.5 Correlation of aging with process variations

During the *Delay modelling* stage, a set of parameters for the *Delay* model was obtained for each (path, process corner) couple. One may think that a different set of parameters for the $\Delta p_{V_{th}}$ model should be computed for each process corner, path and workload triplet. However, the non-correlation between process and aging variations was already demonstrated in [39], as shown in Figure 3.21.

As no correlation is observed between process variation and aging, there is no need to re-estimate the parameters of equation (3.24) for each process corner. An aging simulation has been conducted with both TT and SS process corners to confirm this assumption. The stress condition adopted was $1V$, $125°C$ and 20 years. Table 3.11 shows the threshold voltage shifts for the 5 most degraded

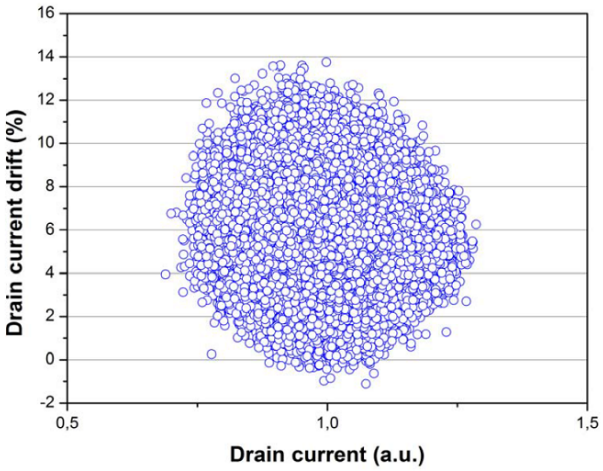


Figure 3.21: Non-correlation between the initial drain current and the drain current drift for more than 1 million devices in $28nm$ FD-SOI technology [39].

transistors in the circuit path. As can be seen, the process corner does not significantly affect the degradation since the difference between the Vth shifts was not larger than 0.05%.

Table 3.11: Threshold voltage shift of the most degraded transistors with typical-typical (TT) and slow-slow (SS) process corners. Stress conditions: $1V$, $125°C$ and 20 years.

Transistor	TT shift (mV)	SS shift (mV)	Difference
# 1	10.3881	10.3928	0.05%
# 2	10.388	10.3927	0.05%
# 3	10.3879	10.3926	0.05%
# 4	10.384	10.3857	0.02%
# 5	10.383	10.385	0.02%

3.5.6 Considerations regarding dynamic variations

The proposed $\Delta p_{V_{th}}$ model has been constructed over SPICE simulations with aging variations using the Eldo UDRM API. As explained in Section 3.3, the API computes the stress stimuli from a transient simulation and then extrapolates the aging degradation for a given power-on time. It is assumed that the circuit endures the same stress during all its lifetime, which is seldom true. Variations of temperature and supply voltage happen constantly during a circuit lifetime, specially when a DVFS strategy is implemented. Therefore, the proposed model must consider these dynamic variations if it is to be used for an on-line estimation.

The first approach proposed to handle dynamic variations in the $\Delta p_{V_{th}}$ model was to take the average degradation into account. In other words, estimate the

$\Delta p_{V_{th}}$ for all considered (V, T) conditions and then compute the weighted average considering the time spent in each condition. Note that SPICE simulators do not allow simulations with variable temperature. However, it is possible to simulate any configuration of supply voltage waveform. The validity of this approach was then verified through a simple V transition. Simulations were conducted with the V at a "low" voltage value for half the time and at a "high" value for the other half. The resulting delay shift was then compared to the average delay shift obtained from simulations with V at "low" and "high" values all the time. Table 3.12 shows the relative delay shifts observed for two cases where the temperature is $125°C$ and the power-on time is 5 years. The delay shifts were measured at 1V. As can be seen, the average degradation is considerably lower than the one obtained with variable V. Moreover, larger the difference between "low" and "high" values of V, larger the error of using an average estimation.

Table 3.12: Delay shifts for two cases of voltage transition, from 0.9V to 1.1V and from 0.8V to 1.2V. The delay shifts were measured at 1V assuming a temperature of $125°C$ and a power-on time of 5 years. The average of the delay shifts for both low and high V is compared to the resulting shift from a variable V, i.e. a simulation where the circuit spends half the time at low V value and the other half at high V value.

Low V		High V		Average	Variable V
0.9V	0.75%	1.1V	2.35%	1.55%	2.07%
0.8V	0.42%	1.2V	3.94%	2.18%	3.30%

From the previous results, we conclude that computing the average degradation is not a suitable approach for handling dynamic variations of supply voltage. To overcome this limitation, a similar method to the one in [124] has been adopted, see Figure 3.22. At a voltage change from V_1 to V_2:

1. $\Delta p_{V_{th}}(V_1, T_x, t_1)$ is calculated, where t_1 is the time spent at V_1.
2. Then, the inverse model is applied to compute t^+, the time spent at V_2 to obtain the same $\Delta p_{V_{th}}$:

$$t^+ = \Delta p_{V_{th}}^{-1} \Leftrightarrow \Delta p_{V_{th}}(V_2, T_x, t^+) = \Delta p_{V_{th}}(V_1, T_x, t_1) \tag{3.25}$$

3. The final $\Delta p_{V_{th}}$ is then computed considering the time spent at V_2 plus t^+:

$$\Delta p_{V_{th}} = \Delta p_{V_{th}}(V_2, T_x, t_2 + t^+) \tag{3.26}$$

We validated this approach for dynamic variations through SPICE simulations with variable voltage. Firstly, we observed that the number of voltage transitions do not change the final delay shift. In other words, the aging degradation is independent of the number of voltage transitions. It depends only on the total

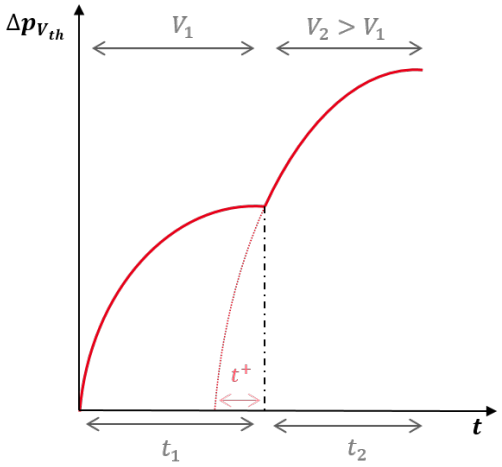


Figure 3.22: Computation of $\Delta p_{V_{th}}$ computation for a voltage transition from V_1 to V_2.

time spent at each voltage level, i.e. the duty cycle. This is consistent with previous works [123] that demonstrated the frequency-independence of BTI. Note that for other aging phenomena such as thermal cycling and electromigration, the rate of voltage or temperature changes is not negligible.

Moreover, the order of voltage levels, i.e. "low" V then "high" V or *vice versa*, does not affect the computed degradation. In fact, the UDRM API extrapolates the degradation of a transient simulation (in nanoseconds) for a given power-on time (in years) taking into account only the percentage of time at each voltage, not their order. In a real experiment, a smaller degradation would be observed in the case where the low supply voltage is used last as a result of the BTI recovery. The reliability models [34] adopted here do incorporate this recovery feature during the time extrapolation therefore avoiding a pessimistic estimation. However, the final degradation reported is the one at the highest voltage value used in the simulation that corresponds to the worst situation. Various voltage scaling situations were simulated with different voltage levels and duty cycles. The path delay was measured at $1V$ and $120°C$ after the stress stimuli were applied. The $\Delta p_{V_{th}}$ was then obtained from the resulting aged delays through equation (3.11). The tested scenarios are listed in Table 3.13 where D stands for the duty cycle of the respective voltage level. Table 3.14 shows the $\Delta p_{V_{th}}$ obtained in simulation and with the proposed approach.

1. $(V_1 = 0.8V, D_1 = 50\%)$, $(V_2 = 1.2V, D_2 = 50\%)$, $t = 10$ years and $T = 120°C$;
2. $(V_1 = 1.0V, D_1 = 50\%)$, $(V_2 = 1.2V, D_2 = 50\%)$, $t = 10$ years and $T = 120°C$;

3. ($V_1 = 0.8V, D_1 = 25\%$), ($V_2 = 1.2V, D_2 = 75\%$), $t = 10$ years and $T = 120°C$;
4. ($V_1 = 0.9V, D_1 = 20\%$), ($V_2 = 1.1V, D_2 = 50\%$), ($V_3 = 1.3V, D_3 = 30\%$), $t = 20$ years and $T = 150°C$;
5. ($V_1 = 0.8V, D_1 = 25\%$), ($V_2 = 1.0V, D_2 = 25\%$), ($V_3 = 1.2V, D_3 = 25\%$), ($V_4 = 1.4V, D_4 = 25\%$), $t = 20$ years and $T = 25°C$;

Table 3.13: Scenarios of variable supply voltage simulated to validated the proposed approach to handle dynamic variations. D_x is the duty cycle respective to V_x.

Scenario	1	2	3	4	5
T	$120°C$	$120°C$	$120°C$	$150°C$	$25°C$
t	10 years	10 years	10 years	20 years	20 years
V_1	0.8V	1.0V	0.8V	0.9V	0.8V
D_1	50%	50%	25%	20%	25%
V_2	1.2V	1.2V	1.2V	1.1V	1.0V
D_2	50%	50%	75%	50%	25%
V_3	-	-	-	1.3V	1.2V
D_3	-	-	-	30%	25%
V_4	-	-	-	-	1.4V
D_4	-	-	-	-	25%

Table 3.14: $\Delta p_{V_{th}}$ obtained in simulation and with the proposed approach for the scenarios defined in Table 3.13.

Scenario	$\Delta p_{V_{th}}$ Simulation	$\Delta p_{V_{th}}$ Model	Difference
1	13.7mV	14.0mV	+2.19%
2	13.7mV	14.0mV	+2.19%
3	15.6mV	15.2mV	-2.56%
4	28.5mV	26.0mV	-8.77%
5	16.6mV	15.8mV	-4.82%

As can be seen, the difference between the simulated $\Delta p_{V_{th}}$ and the modelled one is quite small: even tough an error of 8.77% (2.5mv) is observed in the 4^{th} scenario, it is not significant for the path delay estimation. When this $\Delta p_{V_{th}}$ is integrated into the *Delay* model for a condition of (1V, $125°C$), the computed delay is only 0.56% smaller than the simulated one. These results validate the proposed approach for voltage variations.

Basically, the authors in [124] found that the final degradation experienced by a circuit in a voltage scaling strategy only depends on the time spent at the highest V. Figure 3.23 shows the delay degradation for three DVFS strategies with two voltage levels. The only difference between the three strategies is the value of the

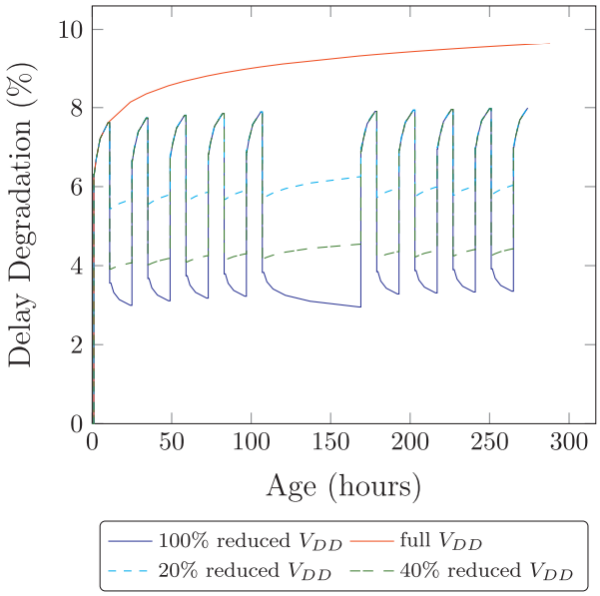


Figure 3.23: Delay degradation curves for different voltage scaling strategies [124]. The time spent at each voltage level as well as the "high" value of V_{DD} is the same for all strategies. The "low" value of V_{DD} changes from 0V to 60% and 80% of the "high" value.

"low" V, which ranges from 0 volts to 60% and 80% of the "high" value. The final delay degradation when a "high" voltage value is applied is identical for the three cases. The same finding is observed for our proposed approach and for SPICE simulations as can be seen in the Table 3.14 for scenarios 1 and 2. Both scenarios produced the same $\Delta p_{V_{th}}$ even tough the "low" voltage level is 0.8V for the first one and 1.0V for the second one. Therefore, it is possible to assume that the circuit is not getting stressed at all when it is at "low" V.

Note that, as previously stated, it is not possible to run SPICE simulations with variable temperature. Therefore, we are not able to validate the adopted methodology for temperature variations as we did for voltage variations. We can only assume that the approach adopted for voltage variations is also valid for temperature variations, taking into account the work performed in [124].

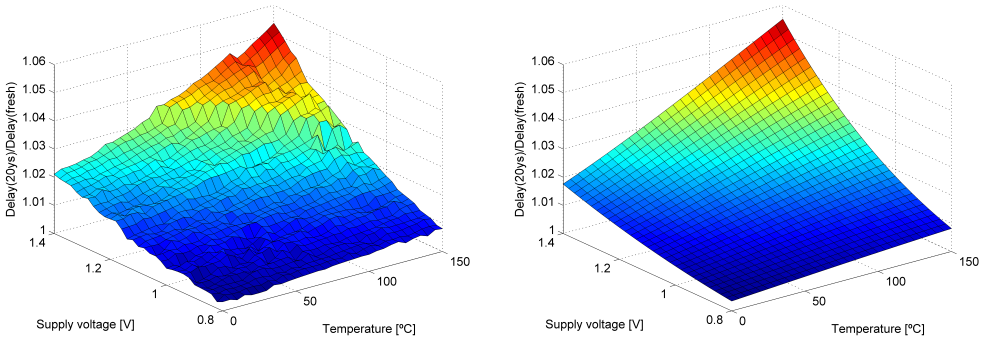
3.6 Validation of the complete model

Once the aging model is constructed, i.e. $\Delta p_{V_{th}}$ model has been tuned, it has to be incorporated into the *Delay* model to assess the degradation of the path delay. With the shift of the parameter $p_{V_{th}}$, the final *Delay* model becomes:

$$Delay(V,T,t) = p_{\beta} + p_{\mu^{-1}}(T)\frac{V}{(V-(p_{V_{th}}(T)+\Delta p_{V_{th}}(\hat{V},\hat{T},t)))^{p_{\alpha}}} \tag{3.27}$$

Note that $\hat{V}$ and $\hat{T}$ in $\Delta p_{V_{th}}$ depend on the historical values of the supply voltage V and the temperature T, respectively. $\hat{V}$ and $\hat{T}$ are equal to the values of V and T used for the *Delay* computation when no dynamic variations are considered. It means that V and T are assumed constant during all the power-on time t.

Figure 3.24 shows a delay shift surface where the same V and T are used for both the $\Delta p_{V_{th}}$ and the *Delay* computations. A power-on time of 20 years is adopted. The delay shift is measured as an increase (%) in the path delay due to the $\Delta p_{V_{th}}$ for the respective (V, T) condition. The degradation surfaces obtained through SPICE simulations and with our proposed models are almost identical. Furthermore, by comparing the path delays simulated in Section 3.5.2 for 28830 (V, T, t) conditions to the respective delays obtained through the *Delay* model, a mean error of $0.15ps$ (0.09%) is observed. The maximum error observed is $0.87ps$ (0.65%) at $(V, T, t) = (1.4V, 150°C, 65 \text{ days})$. The $\Delta p_{V_{th}}$ error at this stress condition was 1.4mV (7.24%). Figure 3.25 gives the cumulative distribution function (CDF) of the delay error. As can be seen, 95% of the errors stayed between ±0.4ps.



(a) SPICE simulations

(b) Proposed model

Figure 3.24: Path delay shift (aged/fresh ratio) where the same (V, T) condition is used to compute both the stress stimuli and the path delay. A power-on time of 20 years is adopted. Left: Spice simulations. Right: Proposed models.

In Figure 3.26, another delay shift surface is generated by taking a constant stress condition of 1.2V, $125°C$ and 20 years for the computation of $\Delta p_{V_{th}}$. Differently from Figure 3.24 where the worst degradation is observed at high voltages due to a increased $\Delta p_{V_{th}}$, in Figure 3.26 it is perceived at low voltages. This is due to a increased sensitivity of the path delay to threshold voltage variations at low voltages ($Delay \propto 1/(V - V_{th})$). The mean error observed was $0.3ps$ (0.15%) while the maximum one was $2.2ps$ (0.68%) at $(V, T) = (0.8V, 0°C)$.

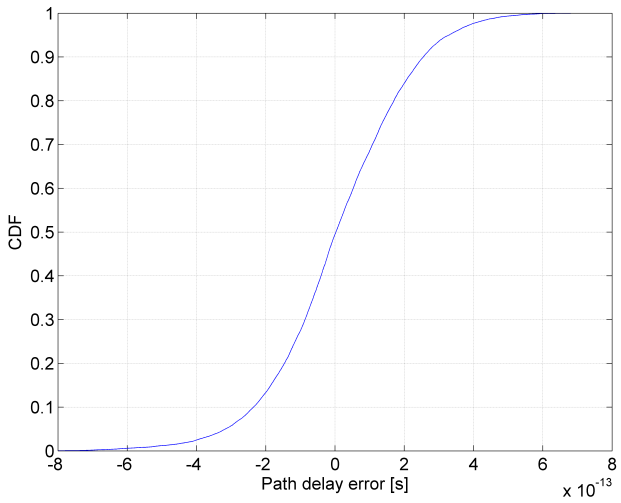
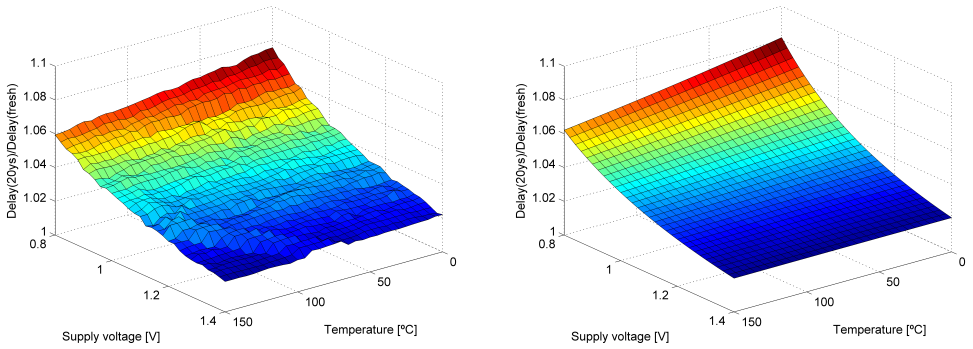


Figure 3.25: Cumulative distribution function of the difference between the simulated delay and the *Delay* model for 28830 (V, T, t) conditions of use. 95% of the errors are between ± 0.4ps.



(a) SPICE simulations

(b) Proposed model

Figure 3.26: Delay shift due to aging (aged/fresh ratio) with a constant stress stimuli obtained at a condition of $(1.2V, 125°C, 20$ years). Left: Spice simulations. Right: Proposed models.

3.6.1 Validation in another benchmark circuit

The proposed methodology has been applied to another benchmark circuit implemented in $28nm$ FD-SOI technology. The methodology is not described in details, as previously done. The objective here is to show that our methodology is circuit independent and remains valid for different architectures. The circuit adopted here consists in a 32-bit Reduced Instruction Set Computer (RISC) Har-

vard architecture, in-order, mono-thread, 5-stage pipeline [134]. Its critical path is composed of 37 cells with a propagation delay of $0.48ns$ at $(V,T) = (1.1V, 75°C)$. The ranges adopted for V, T and t are $[0.8, 1.4]V$, $[0, 150]°C$ and $[0, 20]$ years, respectively, with $\Delta V = 5mV$, $\Delta T = 15°C$ and 10 logarithmically distributed values for t. 1430 (V,T,t) conditions of use were simulated in total. The *Delay* parameters and their respective 95% Confidence Intervals are given in Table 3.15. The fitting resulted in a mean error of $0.38ps$ (0.08%) and a maximum error of $1.17ps$ (0.27%).

Table 3.15: *Delay* model parameters (see equation (3.9)) and respective Confidence Intervals CI for the RISC processor [134].

p_β (95% CI)	2.25e-10 (0.1%)
C_1 (95% CI)	6.96e-11 (0.5%)
k_1 (95% CI)	3.23e-16 (2.0%)
n_1 (95% CI)	2.20 (0.1%)
C_2 (95% CI)	0.41 (0.3%)
k_2 (95% CI)	7.38e-5 (2.2%)
n_2 (95% CI)	1.39 (0.2%)
p_α (95% CI)	3.12 (0.1%)

Table 3.16 presents the $\Delta p_{V_{th}}$ parameters estimated. Note that the wider 95% Confidence Intervals are due to the reduced number of simulated conditions, 20 times smaller than for the DSP circuit. The maximum error was $2.7mV$ at $(V,T,t) = (1.35V, 0°C, 20$ years) while the RMSE was $0.26mV$. Figure 3.27 shows the $\Delta p_{V_{th}}(t)$ curves for both case study architectures and a condition of use of $1.4V$ and $120°C$. As can be seen, the proposed $\Delta p_{V_{th}}$ correctly models the different aging rate depending on the circuit topology.

Table 3.16: $\Delta p_{V_{th}}$ model parameters (see equation (3.24)) and respective Confidence Intervals CI for the RISC processor [134].

γ (95% CI)	4.39 (6.4%)
E_a/k (95% CI)	1129.5 (0.8%)
C_1 (95% CI)	8.77e-3 (9.2%)
C_2 (95% CI)	1.45e-4 (34.2%)
n_1 (95% CI)	0.017 (188.4%)
n_2 (95% CI)	0.216 (5.9%)
a_1 (95% CI)	0.145 (40.0%)
a_2 (95% CI)	0.033 (40.6%)

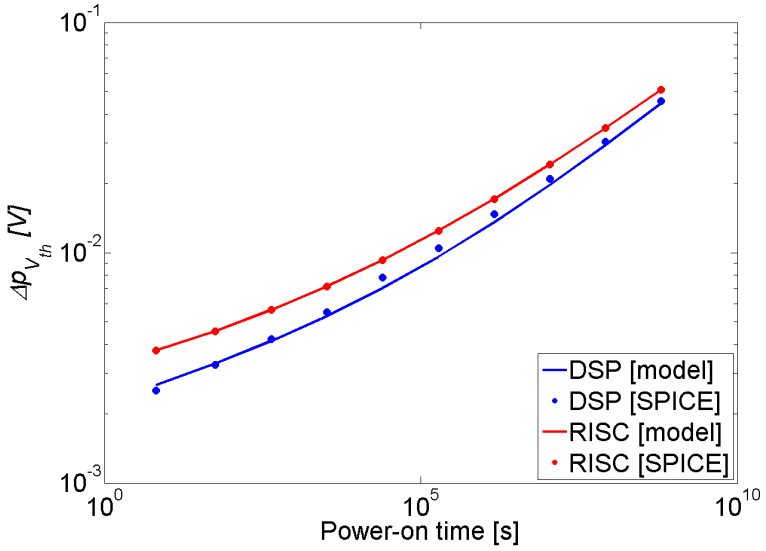


Figure 3.27: $\Delta p_{V_{th}}$ evolution over time for both DSP [61] (blue) and RISC processor [134] (red) obtained through simulations (dots) and the proposed model (lines). The conditions of use are ($1.4V$, $120°C$).

3.7 Conclusion

This chapter proposes a new bottom-up approach for estimating the circuit degradation due to BTI/HCI effects. Built on top of device-level models, it consists in two stages, namely, *Delay Modelling* and *Aging Modelling*. The first stage produces a model, called *Delay*, that models the propagation delay of a circuit path depending on the supply voltage and the temperature. The second one models the aging degradation as a parameter shift Δp of the *Delay* model. The proposed methodology takes into account all factors that impact global aging, namely, circuit topology, workload, supply voltage and temperature variations.

The proposed methodology has been validated on a DSP circuit implemented in $28nm$ FD-SOI technology [61]. SPICE simulations with aging variations were performed with the Eldo UDRM API. A state-of-the-art model coupling both BTI and HCI effects together and featuring BTI relaxation has been adopted [34]. One of the circuit critical paths was simulated within a wide range of supply voltage V and temperature T generating a delay surface. This surface was then fitted to a *Delay* model that was created based on the Sakurai alpha power-law [22]. Fitting results demonstrated the good agreement between the proposed model and the simulated delays, with a mean normalized residual of only 0.08%. Finally, process variations were taken into account in two parameters of the *Delay* model (C_2 and p_α) that depend on the process corner.

During the second stage, an aged delay surface was generated and fitted to

the *Delay* model. By comparing the new parameters with those obtained for the "fresh" surface, a shift of the parameter $p_{V_{th}}$ is observed, $\Delta p_{V_{th}}$. The path was then re-simulated within the same V and T ranges and with a power-on time t up to 20 years. In total, 28830 (V, T, t) conditions were simulated and a different $\Delta p_{V_{th}}$ was obtained for each condition. The surfaces of $\Delta p_{V_{th}}$ exhibited a similar shape than the aging models found in the literature. Therefore, a $\Delta p_{V_{th}}$ model was constructed by testing different configurations of voltage-, temperature- and time-dependence. Fitting the whole $\Delta p_{V_{th}}$ data set with the final model resulted in an average residual of $0.28mV$ with a maximum residual of $4.8mV$. Other simplified versions of the $\Delta p_{V_{th}}$ model are possible but with some accuracy loss.

Finally, both BTI and HCI effects are strongly dependent on the circuit workload, i.e., the signal probabilities and activities. Therefore, new $\Delta p_{V_{th}}$ parameters have to be computed for different sets of signal probabilities. As demonstrated here, some of the parameters can be shared between all workloads to simplify the model without accuracy loss.

Furthermore, the $\Delta p_{V_{th}}$ model is constructed from simulations where a constant voltage V and a constant temperature T are assumed during all the circuit lifetime. An approach similar to the one proposed in [124] was adopted to model dynamic variations. It consists in calculating an "apparent spent time" t^+ whenever a variation occurs. Some scenarios of voltage variation were simulated showing good agreement with the $\Delta p_{V_{th}}$ obtained through this approach.

At the end, the $\Delta p_{V_{th}}$ model was integrated to the *Delay* model. The complete model was validated for the 28830 (V, T, t) conditions previously simulated, resulting in an average error of $0.15ps$ (0.09%) and a maximum error of $0.87ps$. The methodology was also validated on a second benchmark circuit, namely, a RISC processor [134]. Many applications are possible for such an accurate model for aging degradation at circuit-level. For instance, it can be used for building a framework for simulating a system under aging variations as well for on-line estimating the circuit maximum frequency. Some of the possible applications are demonstrated in the following chapter.

The proposed methodology has been validated in $28nm$ FD-SOI technology, nevertheless it is still valid for any technology. The steps described in this chapter would remain the same if the methodology is used with another technology, only the *Delay* and the $\Delta p_{V_{th}}$ model may change. We have demonstrated that both *Delay* and $\Delta p_{V_{th}}$ models fit well for different circuits implemented in the same technology. Therefore, the methodology can be fully automated through scripts for any circuit once the models are built.

Chapter 4

Integration of circuit-level models in different application contexts

In the last years, many works focused on tackling the aging related issues in digital circuits. Some of them [135, 136, 137] consist in predicting the circuit degradation during the design phase. However, BTI and HCI mechanisms depend on the operating conditions, namely, supply voltage, temperature and workload. These conditions are seldom known in the design phase which makes an accurately estimation almost impossible. Other solutions [138, 83, 139] implement slack time sensors in the critical paths to detect the pre-occurrence of setup time violations. Nevertheless, these violations may be produced by a voltage drop or a temperature change; they are not necessarily a consequence of aging. Some other works [111, 140, 109] claim to on-line measuring the aging degradation through ring-oscillator based sensors. Yet, these sensors do not endure the same stress as the functional parts of the circuit and they may degrade in a different rate than the circuit itself.

Therefore, in this chapter we propose the use of the models proposed in Chapter 3 to on-line assess the circuit degradation. They can be used to constantly estimate the change of the circuit maximum operating frequency due to aging and to other dynamic variations. Thus, an adaptation strategy can be implemented to alter the frequency or supply voltage in order to ensure a fault-free operation of the circuit while increasing the energy efficiency. Besides, the models can be applied to on-line estimate the remaining time until circuit failure, i.e. the time when the aging-induced delay shift will exceed a pre-defined safety margin. Moreover, they can be used to determine which are the maximum operating conditions (temperature and supply voltage) at which a circuit can operate without breaking down before a desired lifetime. Finally, the $\Delta p_{v_{th}}$ model can serve as

a reliability parameter to compare identical circuits for instance the processing cores in a multi-core system.

Current BTI and HCI models are mostly device-level models. They can be integrated in SPICE simulators to accurately estimate the aging-induced threshold voltage of each transistor. However, they cannot be used to assess the degradation of a complex circuit, such as a multi-core one. In this chapter, we implement the circuit-level models developed in Chapter 3 to analyze the degradation of a multi-core system. Different task mapping strategies are tested and compared with regard to performance, energy and reliability. The models are also used to simulate the operation of an Adaptive Frequency Scaling (AFS) system. A new method to estimate the degradation of AFS systems is also proposed and evaluated.

This chapter is organized as follows. Firstly, Section 4.1 briefly discusses the state-of-the-art on the abstraction of aging models. Then, different possible uses of the proposed methodology to on-line estimate the circuit degradation are given in Section 4.2. Section 4.3 describes a multi-core simulation framework where both the *Delay* and the $\Delta p_{v_{th}}$ models are integrated to evaluate different task mapping strategies with respect to performance, energy and reliability. Finally, Section 4.4 shows the application of the proposed methodology to simulate a complete AFS system and then it introduces a new technique to on-line assess its aging degradation. The research activities presented in this last section have been published in [10].

4.1 Related works on circuit-level aging modelling

Previous works have already focused on abstracting the complexity of existing device-level aging models, in particular NBTI ones. Some works managed to abstract the complexity of aging models from device-level to gate-level. For instance, [141] generated gate-level models for NBTI degradation while [142] modelled both BTI and HCI effects at gate-level. Nevertheless, a critical path in complex circuits may have 50 or more gates. Although such models simplify aging simulations, they are still not suitable to be used for on-line estimation due to the required computational complexity.

In [143], a modeling framework is proposed for timing variations in a RISC processor. A second order polynomial is used to model the path delay dependence on temperature. However, instead of integrating the supply voltage in their model, the authors obtain a set of model parameters for each value of the supply voltage. This allows the estimation of the temperature-induced f_{MAX} shift for different supply voltages, but not the estimation of the shift due to voltage variations. Moreover, the authors do not model the NTBI degradation in the path delay. Finally, only the typical process corner is addressed.

In [144], the authors modelled the workload-dependence of NBTI. Their methodology consists in an off-line modelling stage and an on-line monitoring stage. First, a method is proposed to select a small set of the most representative flip-flops in a circuit with regard to aging. For each possible workload, the signal probabilities (SPs) of the selected flip-flops are obtained through cycle accurate RTL simulations. Then, the aging-induced delay shift of the critical path is estimated for each workload through BTI-aware timing analysis. An aging model is then constructed correlating the SPs of the selected flip-flops to the resulting delay increase. During run-time, the SPs of the selected flip-flops are monitored through counters attached to them. From the obtained signal probabilities, the aging-induced delay shift is then calculated using the aging model constructed at design phase. However, this work only addresses the workload-dependence, considering both the temperature and the supply voltage as constant.

A methodology very similar to the one we propose in the present work was recently published in [145]. The methodology is also based on aging-aware simulation and data fitting with simulated path delays through MATLAB. Models are generated for the NBTI-induced delay shift of critical paths. The resulting models have an exponential dependence on power-on time. They are validated on different architectures and workloads, where two sets of model parameters are obtained for each architecture. The first set of parameters corresponds to the lower f_{MAX} bound and the second one to the higher f_{MAX} bound. They are obtained from the worst and best workloads with regard to aging degradation, respectively. However, such models are built for a unique Process-Voltage-Temperature (PVT) condition. Therefore, nor the voltage variations neither temperature variations

are taken into account.

Therefore, and to the best of our knowledge, this PhD thesis is the first work to take into account all sources of variation (PVT and aging effects) in behavioral circuit-level models.

4.2 On-line estimation of circuit degradation

Current solutions to measure the degradation of a circuit are mainly based on ring-oscillators [111, 140, 109]. These sensors are composed of two identical units. One of the units is constantly stressed while the other one is kept off with power gating. It is possible to assess how much the stressed unit was degraded by comparing their oscillating frequencies. However, these canary structures do not degrade at the same rate than the functional parts of the circuit. Firstly, because they do not own the same topology as the circuit critical paths and, secondly, aging highly depends on the stressing input patterns which are determined by the application running in the circuit and the input data. These conditions cannot be reproduced on ring oscillators.

In this section, we present four possible uses of the proposed methodology to accurately estimate the degradation of a circuit during its operation. Dynamically tracking the aging-induced delay shift allows the establishment of strategies to increase the circuit reliability as well as its energy efficiency. The implementation costs are not discussed here because different ways to compute both the *Delay* and the $\Delta p_{V_{th}}$ models can be implemented. Note that their implementation will be most likely done as an application software. However, it is possible to design a dedicated hardware for it when the area and power overheads are compensated by the gains offered by the proposed methodology.

4.2.1 Control loop for Adaptive Voltage and Frequency Scaling

As stated in Chapter 1, the circuits implemented in advanced CMOS nodes are highly sensitive to variability. Large voltage margins must be then added to ensure a reliable operation of the circuit as shown in Figure 4.1. Otherwise, "sense & react" approaches can be implemented to increase the energy efficiency by removing such margins. Those techniques rely upon embedded sensors that are placed in the circuit critical paths and monitor the change in the slack time [138, 83, 139].

However, the implementation of such sensors considerably increases the circuit design complexity. The sensors alter the timing characteristics of the critical paths resulting in additional timing optimization steps. Furthermore, this is a reactive technique since the circuit adapts itself only after the delay shift occurs. A proactive solution, which foresees the delay shift before it occurs, would be a better choice for avoiding irreversible variations such as aging-induced ones.

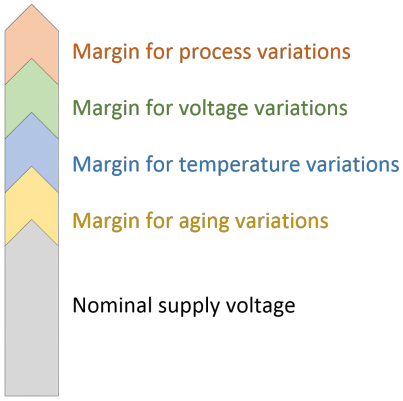


Figure 4.1: Large voltage margins are required due to the increased variability in advanced technology nodes.

The methodology proposed in Chapter 3 can be adopted as an alternative solution to the traditional slack time sensors. Instead of using invasive monitors, the proposed methodology only requires supply voltage and temperature monitoring. Moreover, it allows to predict the effect of any kind of variation (supply voltage, temperature, aging) on the circuit maximum operating frequency f_{MAX}. Figure 4.2 shows the diagram flow of the envisioned solution.

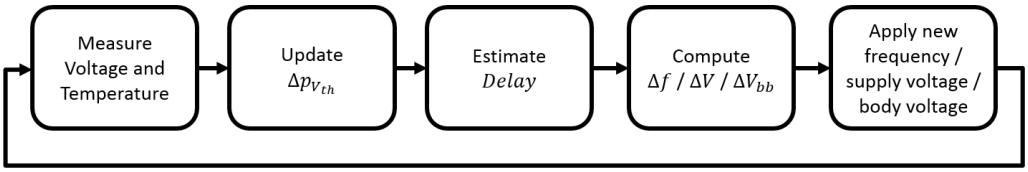


Figure 4.2: Closed-loop strategy (i.e. sense and react) to reduce energy consumption by dynamically updating the circuit frequency, supply voltage or body voltage based on the estimated *Delay*.

Its implementation consists in a closed-loop scheme where embedded monitors periodically provide measures of the supply voltage and of the circuit temperature. The aging-induced parameter shift $\Delta p_{V_{th}}$ is estimated only when it seems necessary. Note that circuit aging has a power dependence on time. Therefore $\Delta p_{V_{th}}$ must be estimated more often at the beginning of the circuit lifetime than after some months of operation. The *Delay* may be periodically estimated or whenever a significant change of the supply voltage or of the temperature is observed. Finally, the circuit adapts itself based on the estimated *Delay*. This depends on the adaptation strategy implemented in the circuit, for instance the adaptation can be done through the clock frequency, the supply voltage or even the body voltage.

We now illustrate the benefits of using an Adaptive Voltage Scaling (AVS)

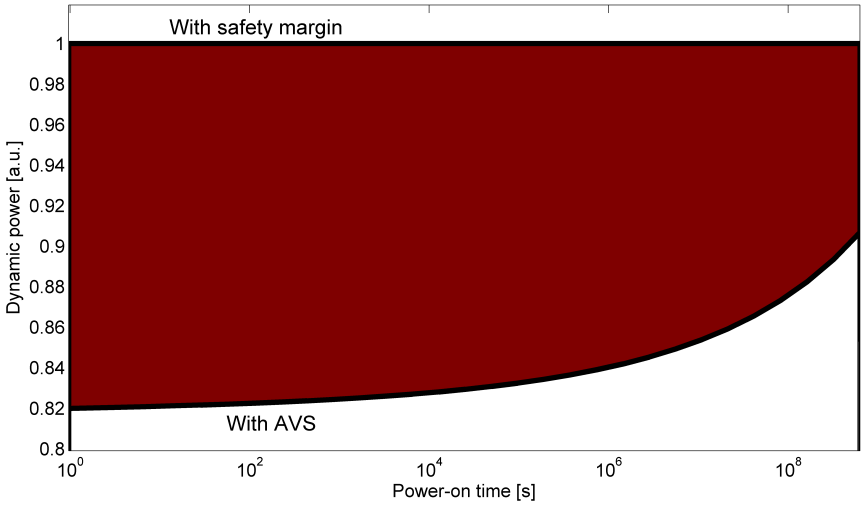


Figure 4.3: Comparison of the dynamic power dissipated with a safety margin and with Adaptive Voltage Scaling (AVS). The red area corresponds to the total energy reduced which is equivalent to 11.1% after 20 years.

strategy to mitigate aging variations. We estimate the dynamic power dissipated by a circuit for two cases, namely when a supply voltage guard-band is used and the when an AVS strategy is implemented instead. The model parameters computed in Section 3.6.1 for a RISC processor [134] are adopted. At the beginning of its lifetime, this RISC processor requires a supply voltage of $1.35V$ to operate at a clock frequency of 2.5GHz for a temperature of $120°C$. However, the supply voltage necessary to ensure its correct operation at a worst case temperature of $150°C$ and for a power-on time of 20 years is equal to $1.48V$. This $130mV$ guard-band results in an increase of 25.8% of the dynamic power dissipated at the beginning of circuit lifetime. Figure 4.3 shows the energy gain by using AVS instead of a safety margin. The total energy consumed after 20 years is reduced by 11.1%.

The computation of equation (3.10) is a complex operation. Such computation requires tens or hundreds of clock cycles if it is performed at software-level. It could be computed quite faster if a dedicated hardware is designed for it. However, it would result in a considerable area and power overhead making it a bad trade off. If the *Delay* cannot be updated in a few clock cycles, the circuit will not be able to adapt against fast variations such as supply voltage variations. In this case, the slack time sensors would be necessary to mitigate fast variations.

Nevertheless, another possible implementation of the proposed models in an adaptive system is by using a Look-Up Table (LUT). The circuit f_{MAX} (inverse of *Delay*) is initially computed for a given condition of use (V, T). A LUT is then constructed with the resulting f_{MAX} shifts for small variations of V and T.

Figure 4.4 shows an example of LUT for $(V, T) = (0.8V, 100°C)$ constructed with the *Delay* model parameters computed for the RISC processor of Section 3.6.1. The circuit can operate at a clock frequency of 1.160 GHz at the nominal condition. For instance, if a voltage drop of 4mV occurs, the clock frequency must be then reduced by 11 MHz to avoid timing faults. The LUT is updated whenever a significant change of $\Delta p_{V_{th}}$ is computed. In a DVFS system it is possible to have one LUT for each voltage level. This solution allows a fast adaptation (few clock cycles), in presence of small variations without the need of additional circuitry.

T \ V	-6mV	-4mV	-2mV	0.8V	+2mV	+4mV	+6mV
-15ºC	-29MHz	-24MHz	-18MHz	-12MHz	-6MHz	+0MHz	+6MHz
-10ºC	-25MHz	-20MHz	-14MHz	-8MHz	-2MHz	+4MHz	+9MHz
-5ºC	-21MHz	-15MHz	-10MHz	-4MHz	+2MHz	+8MHz	+13MHz
100ºC	-17MHz	-11MHz	-6MHz	**1.160 GHz**	+6MHz	+11MHz	+17MHz
+5ºC	-13MHz	-7MHz	-2MHz	+4MHz	+10MHz	+15MHz	+21MHz
+10ºC	-9MHz	-3MHz	+2MHz	+8MHz	+14MHz	+19MHz	+25MHz
+15ºC	-5MHz	+1MHz	+7MHz	+12MHz	+18MHz	+23MHz	+29MHz

Figure 4.4: Example of LUT table for an AFS system constructed with the models developed in Chapter 3. It contains the induced shift of f_{MAX} for small variations of the supply voltage and the temperature.

4.2.2 Dynamic Mean Time to Failure (MTTF) computation

Safety margins against aging variations have to be considered in a circuit when no adaptation strategy is implemented. In other words, the clock frequency applied to the circuit is lower than its f_{MAX} to ensure that it can properly work even after some years of operation. The Mean Time to Failure (MTTF) is measured at the presilicon design phase as the expected time that the aging-induced delay shift is going to exceed the safety margin $Delay_{Margin}$ for a given condition of use (V, T). However, the condition of use is seldom constant over the circuit lifetime. The same stands for the workload which affects as well the circuit aging. The MTTF therefore cannot be correctly estimated at the design phase.

The proposed methodology can be used to calculate the MTTF during the circuit operation based on the actual conditions of use. Using the historical average values of voltage, temperature and workload, one only needs to apply the inverse function of $\Delta p_{V_{th}}$ to find MTTF. This procedure is quite similar to the one adopted for dynamic variations in Section 3.5.6. Using the inverse of

equation (3.10), one can find the value of $\Delta p'_{V_{th}}$ corresponding to a given margin $Delay_{Margin}$ for an average condition of use $(\hat{V}, \hat{T})$:

$$\Delta p'_{V_{th}} \mid Delay(\hat{V}, \hat{T}, \Delta p'_{V_{th}}) = Delay_{Margin} + Delay(\hat{V}, \hat{T}, 0) \tag{4.1}$$

The inverse of equation (3.24) is then applied to find the time necessary to reach this $\Delta p'_{V_{th}}$:

$$t_{MTTF} \mid \Delta p_{V_{th}}(\hat{V}, \hat{T}, t_{MTTF}) = \Delta p'_{V_{th}} \tag{4.2}$$

The t_{MTTF} corresponds to the overall MTTF without taking into account the time already spent t_{spent}. If there was no significant change in the condition of use during the circuit operation, the remaining lifetime t_{remain} is then simply given by:

$$t_{remain} = t_{MTTF} - t_{spent} \tag{4.3}$$

Otherwise, an equivalent spent time t^+ must be calculated by applying again the inverse of equation (3.24) but now on the current aging-induced parameter shift $\Delta p_{V_{th}\,current}$:

$$t^+ \mid \Delta p_{V_{th}}(\hat{V}, \hat{T}, t^+) = \Delta p_{V_{th}\,current} \tag{4.4}$$

As a consequence:

$$t_{remain} = t_{MTTF} - t^+ \tag{4.5}$$

By dynamically estimating the MTTF, it is then possible to know how close the circuit is to get in an unreliable state. When the aging-induced delay shift becomes larger than the initial $Delay_{Margin}$, either the circuit must be stopped or error detection techniques must be implemented.

4.2.3 Maximum operating conditions

In some situations, it is essential to ensure that a circuit will last a certain time before getting in an unreliable state. This is the case for avionics and space applications where the embedded systems must endure at least 20 years of operation in a harsh environment [6]. As the aging rate highly depends on the conditions of use (V, T) and the workload, this constraint may not be always fulfilled. Therefore, it is important to evaluate the worst conditions at which a circuit can run while guaranteeing that the lifetime constraint will be respected.

The proposed models can be then used to calculate the maximum operating conditions possible so that the aging-induced parameter shift will not exceed a given safety margin $\Delta p'_{V_{th}}$ before a given lifetime t_{end}. Average values can be

adopted for the parameters in equation (3.24). Otherwise, the values corresponding to the worst case workload can be used. The objective may be to find the maximum temperature T_M for a given supply voltage V_x:

$$T_M \mid \Delta p_{V_{th}}(V_x, T_M, t_{end}) = \Delta p'_{V_{th}} \tag{4.6}$$

or to find the maximum supply voltage V_M for a given temperature T_x:

$$V_M \mid \Delta p_{V_{th}}(V_M, T_x, t_{end}) = \Delta p'_{V_{th}} \tag{4.7}$$

Moreover, it is possible to construct a pareto frontier for the maximum values of V and T as shown in Figure 4.5. Note that the circuit temperature is dependent on the power dissipated and, consequently, on the supply voltage. Therefore, an increase in the supply voltage will inevitably raise the temperature.

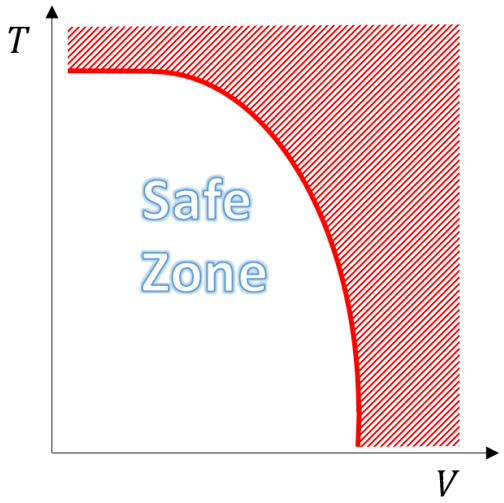


Figure 4.5: Pareto frontier of maximum values for the supply voltage V and the temperature T for reaching a given lifetime t_{end} without the aging-induced parameter shift exceeding a given safety margin $\Delta p'_{V_{th}}$.

Obviously, the conditions of use V and T are seldom constant. They evolve during the whole circuit lifetime. So, the maximum operating conditions must be dynamically updated considering the current $\Delta p_{V_{th}}$. Finally, whenever V and T are higher than the maximum allowed ones, an action has to be taken to lower them and get the circuit back in the *Safe Zone*.

4.2.4 Simplified on-line estimations

The methodology proposed in Chapter 3 takes into account the effect of all sources of variations on the circuit path delay. However, in some situations a simplified implementation may be preferred. For example, consider a system where both

the supply voltage V and the temperature T are supposed to be constant. In this case, the *Delay* is (dynamically) only affected by aging variations. This means that it can be directly inferred from $\Delta p_{V_{th}}$ and that there is no need to compute equation (3.10). To ease the process, a LUT can be created in advance with entries of the induced delay shift for different values of $\Delta p_{V_{th}}$ ($[\Delta p_{V_{th}} \rightarrow Delay]$ or $[\Delta p_{V_{th}} \rightarrow f_{MAX}]$).

Besides that, with V and T constant, the computation of $\Delta p_{V_{th}}$ is also simplified. Basically, it depends only on the power-on time and on the workload. If no significant difference is observed between the impact of all possible workloads on the resulting aging, it can even be directly determined from the power-on time. Again, a table can be constructed with the $\Delta p_{V_{th}}$ values for some power-on times.

Finally, in a many-core circuit where tens or hundreds of twin processors are integrated, the $\Delta p_{V_{th}}$ can be directly used to compare the degradation endured by each core. Since all cores have the same critical paths, there is no need to compute the aging-induced delay shift of each core. The idea is to use the measure of $\Delta p_{V_{th}}$ in task mapping strategies that favor the less degraded cores over the more degraded ones. As a consequence, all the cores will degrade at the same rate and avoid that some of them will become slower than the others. This idea is demonstrated with more details hereafter.

4.3 Aging-aware task mapping in multi-core context

Parallelism was the key element to continue increasing the processors performance after the "power wall" was hit due to significant heat dissipation. That raised the development of multi-core processors in the last decade and of the so-called many-core processors, with tens or hundreds of cores. Some examples of many-core architectures are the Intel's Xeon Phi series, with up to 72 cores, and the Kalray's MPPA2-256, composed of 288 cores. Considerable efforts have been done to get the most from these architectures, specially on the improvement of task mapping strategies. The objective of these strategies is to achieve load balancing that gives the best trade-off between performance and energy consumption.

However, reliability has become a major issue in advanced technology nodes with the emergence of new challenges in the use of multi-core processors. Besides performance and energy, it is now essential to also take aging degradation into account during task mapping. Otherwise, an unbalanced aging of the whole processor may appear leading to a premature wear-out of some processing units. Moreover, the increase in power density resulted in the emergence of dark silicon. This latter corresponds to the part of a circuit that cannot be powered-on due to thermal design power constraint. Recent researches claimed that the amount of dark silicon may reach up to 80% of the circuit with $8nm$ CMOS technology [25]. Thus, having a measurement of the aging degradation of each core would

help keeping the fresher cores powered-on instead of the more aged ones in a dark silicon context.

Previous works focused on solutions for task mapping in multi-core processors targeting a better balance between performance, power and reliability. Most of them are governed by thermal related mechanisms such as thermal cycling [146]. These works basically focus on reducing the circuit temperature which is considered to be linked to the circuit the reliability. Some other works focus on BTI and HCI effects [147, 148], but they employ generic and not fully appropriate models to assess the reliability of the circuit. In this subsection we propose a multi-core simulation framework to assess the impact of different task mapping strategies on the circuit reliability. The models developed in Chapter 3 allow to accurately measure the degradation endured by each core for the different strategies. The objective is to check if the strategy which results in the best trade-off between performance and energy is also the one which leads to the smaller circuit degradation.

4.3.1 Description of the multi-core simulation framework

The multi-core simulation framework is implemented in the Matlab environment. It is basically composed of three elements, namely, *Core*, *Task* and *Scheduler*. At the beginning of a simulation, n *Cores* are instantiated. A *Scheduler* then dynamically allocates a list of *Tasks* between the instantiated *Cores* based on a chosen task mapping strategy.

The objective for this multi-core simulation framework is to evaluate how the proposed models could be used to compare different task mapping strategies with respect to performance, energy and reliability. The performance is defined by the ratio of *Tasks* completed before and after their respective deadline. The energy accounts for the energy consumed by all *Cores* considering dynamic and static powers. Finally, the reliability is measured by the value of $\Delta p_{V_{th}}$ for each *Core*.

A *Task* is defined by:

- the initial time t_{ini}, which is the time when the *Task* is created;
- the deadline t_{end}, which corresponds to the time when the *Task* must be completed;
- the load δ, estimated as the number of processing cycles required to complete the *Task*;
- the activity factor α, used to compute the dynamic power dissipated by the *Core* during the *Task* execution;
- the parameters in the $\Delta p_{V_{th}}$ formula, used for estimating the aging degradation endured by the processing *Core*.

The minimum clock frequency f_{min} required to complete a *Task* before its deadline is defined as follows:

$$f_{min} = \frac{\delta}{(t_{end} - t_{ini})} \tag{4.8}$$

which is valid only if the *Task* starts to run immediately, i.e. if there is a *Core* available when the *Task* is created. Otherwise a higher clock frequency would be necessary.

A *Task* can be executed by only one processing *Core*. It is not allowed to assign its execution to more than one *Core*. Moreover, a *Core* cannot starts another *Task* before finishing the one that is already in execution. Performance overhead due to context switching is therefore not taken into account. A *Core* can execute only one *Task* at a time, still it has its own waiting *Tasks* list. When the current *Task* is finished, the *Core* instantly starts executing the first *Task* in its waiting list, if it is not empty.

The *Scheduler* implements a task mapping strategy to assign each newly created *Task* to a *Core*. If the chosen *Core* is already executing another *Task*, the new one is put in its *Task* waiting list. Each *Core* has a field with the sum of the loads δ of all *Tasks* assigned to it. This information may be used by the *Scheduler* in its task mapping strategy. The other *Cores* data that the *Scheduler* can access is their $\Delta p_{V_{th}}$, which can be used by the mapping strategy to favor the fresher *Cores* over the more degraded ones. The *Scheduler* is also in charge of setting the clock frequency f_{clk} of each *Core*. We suppose that 11 frequency levels are available, from 1.5GHz to 2.5Ghz with a frequency step of 100MHz. The concept of the many-core simulation framework is shown in Figure 4.6. It depicts the data fields of both *Tasks* and *Cores* that are visible to the *Scheduler*.

We consider that an AVS strategy is implemented in each *Core*. In other words, the supply voltage V applied is always the minimum necessary one to assure a fault free operation. V is calculated as an inverse function of $Delay$ (see equation (3.9)) considering the clock frequency applied ($Delay = 1/f$). The RISC processor [134] presented in Section 3.6 is used as the benchmark architecture here. All *Cores* are instantiated with the parameters shown in Table 3.15. Process variations were not included, which could be possible without extra difficulty. The supply voltages necessary for a "fresh" *Core* to run at 1.5GHz and 2.5GHz are 0.92V and 1.33V, respectively, for a temperature of $75°C$.

An ambient temperature is set for all *Cores*. Nevertheless, their junction temperatures depend on their respective dynamic powers. The power dissipated by the circuit path is obtained from SPICE simulations. Simulations were conducted within a wide range of supply voltage and temperature generating a surface for the dynamic and the static power. Both surfaces were then fitted to polynomial equations through regression analysis. A quadratic equation is adopted for the

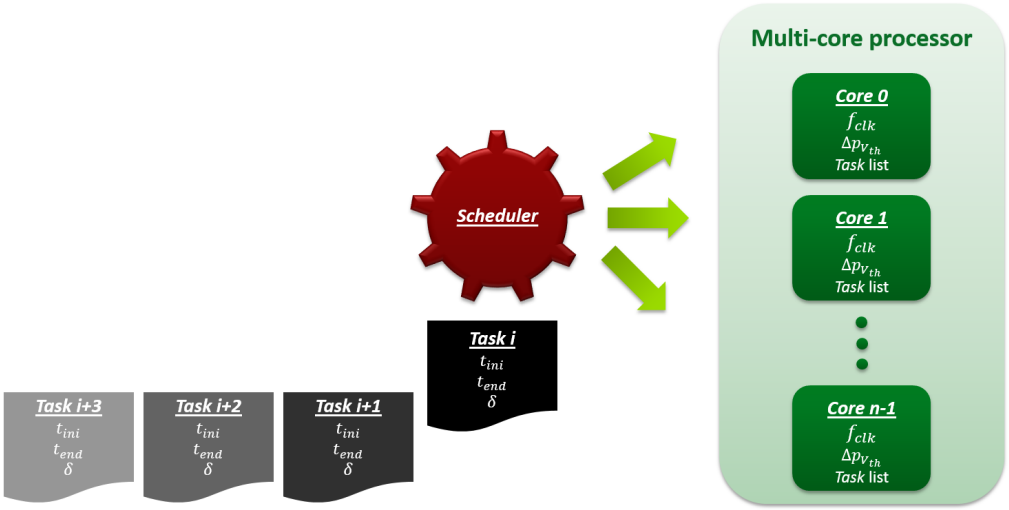


Figure 4.6: The multi-core simulation framework with its 3 elements, namely *Tasks*, *Scheduler* and *Cores*. The *Scheduler* has access to the displayed data fields of both *Tasks* and *Cores* in order to perform the task mapping.

dynamic power since it depends mainly on the supply voltage:

$$P_{dyn}(V) = C_0 + C_1 V + C_2 V^2 \tag{4.9}$$

For the static power, a polynomial equation with a degree of 4 for the supply voltage and a degree of 3 for the temperature is used:

$$\begin{aligned} P_{stat}(V,T) =& C_0 + C_1 V + C_2 T + C_3 V^2 + C_4 VT + C_5 T^2 + C_6 V^3 + C_7 V^2 T + \\ & C_8 V T^2 + C_9 T^3 + C_{10} V^4 + C_{11} V^3 T + C_{12} V^2 T^2 + C_{13} V T^3 \end{aligned} \tag{4.10}$$

As the objective is only to compare the task mapping strategies, a normalized total power is used instead of an absolute one. The power is normalized to 1.1V, 75°C, 2GHz and an activity factor of 0.35.

A script generates the list of *Tasks* based on some parameters, such as the simulation end time and the number of *Cores*. The total number of *Tasks* depends on a "density" parameter. For a density equal to 1, the number of simultaneous *Tasks*, in average, will be equivalent to the number of instantiated *Cores*. For a density of 2, it will be twice as many *Cores*. Another parameter is the maximum frequency f_{min} required to finish a *Task* on time. The load δ and the duration (t_{end}−t_{ini}) of each *Task* are randomly chosen but always respecting this maximum f_{min} (see equation (4.8)). The total number of *Tasks* and their average duration time are correlated and they depend on the simulation end time and on the "density" parameter. They can be thus defined in two ways. The number of *Tasks* can be directly chosen, so the average duration time of each *Task* is determined

based on it. Inversely, the average duration time can be specified and then the number of *Tasks* is determined from it.

Finally, either a random activity factor α is assigned to each *Task* or a constant activity is defined for all of them. α is used to calculate the dynamic power dissipated by the processing *Core* during the *Task* execution. The same goes for the $\Delta p_{V_{th}}$ parameters. A single set of parameters can be used for all *Tasks* or different sets are randomly allocated between the *Tasks*.

4.3.2 Task mapping strategies: Performance × Energy × Reliability trade-off

The multi-core simulation framework described in the previous subsection is now used to evaluate different task mapping strategies. A circuit with 8 *Cores* is simulated. All *Cores* are assumed to be identical. The list of *Tasks* used as input was generated using the following parameters:

- 20000 *Tasks* in total;
- Simulation end time equal to 10 years;
- Average *Task* duration time around 35 hours.
- The minimum and the maximum *Task* duration time are 1/4 and 7/4 of the average one (around 9 and 61 hours), respectively;
- A "density" parameter equal to 1;
- A maximum f_{min} of 2GHz, i.e. all *Tasks* are created requiring a clock frequency up to 2GHz to be completed before deadline;
- A constant activity α equal to 0.2 for the dynamic power estimation;
- The set of $\Delta p_{V_{th}}$ parameters shown in Table 3.16 are adopted for all *Tasks*.

A simulation with this input list takes between 12 and 15 minutes to be completed on a quad-core Intel i5 processor operating at 3.20 GHz. The simulation time depends on the *Scheduler* used. The different task mapping strategies tested here are:

- *Scheduler 0* distributes the *Tasks* in a cyclic way. It starts from *Core 0* until *Core 7*, repeatedly. It implements DVFS by appling the minimum clock frequency necessary to complete the new *Task* before its deadline. The sum of all other *Tasks*'s load δ in the *Core*'s waiting list is also taken into account in the computation of this minimum frequency;

- *Scheduler 1* chooses the "less charged" core, i.e. the one with less remaining cycles to be computed considering all *Tasks* in its waiting list. If there is more than one core in this condition (normally when no task is being executed), the one with smaller $\Delta p_{V_{th}}$ is chosen. It also implements DVFS as *Scheduler 0*;

- *Scheduler 2* is similar to *Scheduler 1* but without DVFS. A constant clock frequency of 2 GHz is adopted;

- *Scheduler 3* is similar to *Scheduler 1*. However, when there are more than 3 *Cores* busy, it increases the clock frequency of all *Cores* that are running at a low frequency. If there are 4 *Cores* busy, it assures that all of them operate at 1.6GHz at least. This minimum frequency is gradually increased up to 2GHz for when all the 8 *Cores* are busy.

- *Scheduler 4* is similar to *Scheduler 2* except that the maximum clock frequency possible, i.e. 2.5 GHz, is applied to all cores;

The simulation results are given in Table 4.1 as the average values for all 8 *Cores*. The performance is measured as the number of delayed tasks. The energy dissipated is calculated from a power normalized to 1.1V, 75°C, 2 GHz and α equal to 0.35. Finally, the estimated $\Delta p_{V_{th}}$ is used as the reliability measurement. Note that a lower value represents a better result for the 3 indicators. The last column gives then the inverse of the product of the 3 indicators which can be used as an overall efficiency indicator for each *Scheduler*.

Table 4.1: Performance, energy and reliability measures for each *Scheduler*. Performance is the amount of delayed tasks, energy is calculated from a normalized power (w.r.t. 1.1V, 75°C, 2 GHz and α equal to 0.35) and reliability is the estimated $\Delta p_{V_{th}}$. The last column gives the inverse of their product as an overall efficiency indicator.

Strategy	DVFS	Performance	Energy	$\Delta p_{V_{th}}$ [mV]	$1/(P \times E \times \Delta p_{V_{th}})$
Sched 0	Yes	233.12	86.80	15.86	3.12e-6
Sched 1	Yes	103.62	76.47	11.99	10.53e-6
Sched 2	No	196.87	93.69	8.10	6.69e-6
Sched 3	Yes	87.62	85.37	10.76	12.43e-6
Sched 4	No	75.12	160.34	24.26	3.42e-6

Remind that *Scheduler 0* actually does not implement any task mapping strategy, it only distributes the *Tasks* between the *Cores* in a cyclic way. On the other hand, *Scheduler 2* implements a task mapping strategy but it does not apply DVFS. By comparing the results of both strategies, *Scheduler 2* reduces the number of delayed tasks by 16% and $\Delta p_{V_{th}}$ by 49% with a slight energy increase equal to 8%. From these results, it can be seen that a correct load-balancing

strategy leads to a considerable improvement in circuit reliability. Yet, better figures are reached by applying DVFS together with a task mapping strategy, as done by *Scheduler 1*. Despite a small $\Delta p_{V_{th}}$ increase, its overall efficiency is 57% better compared to *Scheduler 2* (last column in Table 4.1).

Scheduler 3 manages to produce even better results by only increasing the clock frequency of the *Cores* operating at low frequencies when the circuit is getting overloaded with *Tasks*. Besides reducing the amount of delayed tasks, this strategy also reduces $\Delta p_{V_{th}}$ by minimizing the need of applying very high frequencies and, consequently, very high voltages. Indeed, as discussed in Section 3.5.6, the low value of voltage does not significantly impact the aging degradation, only a high value of voltage and the time spent at it impact the core degradation. The supply voltage should always be at its minimum necessary value in an ideal scenario regarding reliability. This is why *Scheduler 2* resulted in the lowest value for $\Delta p_{V_{th}}$ besides using the medium value for the clock frequency.

Figure 4.7 gives a visual representation of the results presented in Table 4.1. Each indicator has been inverted and then normalized to its respective highest value. As it can be seen, *Scheduler 4* is the best strategy possible with regard to performance. On the other hand, it is the worst one with regard to reliability and energy consumption. Finally, while *Scheduler 3* is not the best in any aspect, it is the one who gives the best trade-off between the three indicators.

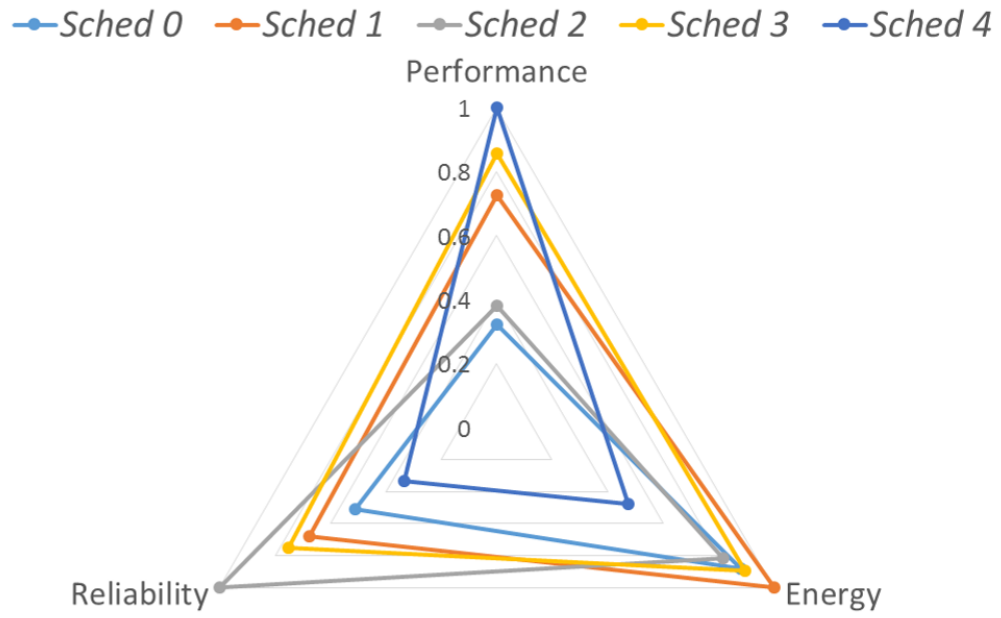


Figure 4.7: Indicators of performance, energy and reliability inverted and normalized to their respective highest values.

4.4 Simulation of an AFS system under aging variations

In Section 2.2 we introduced the concept of adaptive architectures as a solution for coping with variability in digital circuits. These circuits adapt themselves on the fly to avoid timing faults while increasing the energy efficiency. For instance, an Adaptive Frequency Scaling (AFS) system is composed of a variable frequency generator and a closed-loop control to modify the clock frequency depending on the outputs of some embedded monitors, for instance some slack time sensors [138, 139, 83]. The sensors are inserted in the circuit critical paths. Then, they raise a warning when the path delay is close to the clock period.

Yet, these monitors can only sense the variation in the path delay, they do not detect the origin of the variation. A reduction of the circuit maximum operating frequency (f_{MAX}) might be due to a simple voltage drop or a change of the temperature. Nevertheless, it might as well be the result of the circuit degradation. In this subsection we propose a method to measure the aging-induced performance shift of an AFS system by tracking both f_{MAX} and the temperature evolution over time. This technique uses sensors present in any adaptive architecture and it does not require additional structures.

In order to assess the feasibility of the proposed solution, we must simulate an AFS architecture with its embedded sensors under aging variations. It is not feasible through SPICE simulation due to the excessive computational time required to simulate every critical path in various use conditions. Therefore, a platform that features circuit-level aging models must be implemented. That can be done using the methodology proposed in Chapter 3. The circuit critical paths are modeled and the resulting models are then integrated in the simulation platform. This allows the simulation of an adaptive system working under all sources of variability, namely, process, voltage, temperature and aging.

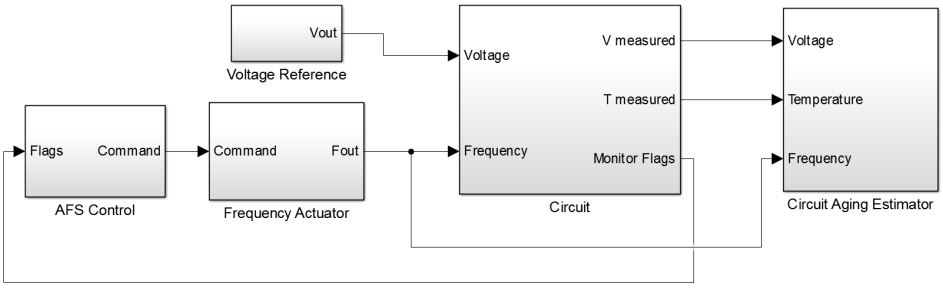
4.4.1 Complete AFS system model in Simulink

The DSP architecture described in Section 3.3.1 was adopted as benchmark. 80 of its critical paths were modeled. The generated models ($Delay$ and $\Delta p_{V_{th}}$ for each path) were then integrated in the MATLAB/Simulink environment representing the in-situ delay monitors. A warning is raised whenever a modelled path delay is equal to the clock period. Previous works demonstrated the correlation between the circuit aging and the slack monitor aging [83]. In other words, the circuit maximum operating frequency f_{MAX} and the monitor warning frequency degrade both at the same rate.

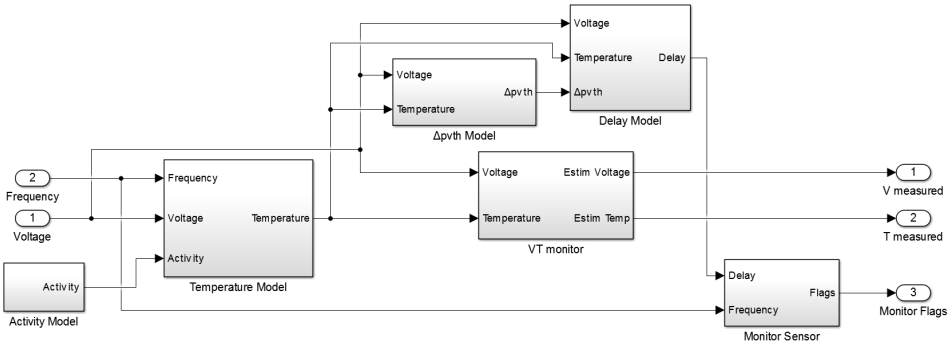
Note that the occurrence of warning flags highly depends on the circuit workload [149]. If a monitored path is not being stimulated, it will not raise a warning even tough its propagation delay is close to or larger than the clock period. There must be a signal transition at the path end so that the pre-occurrence of a tim-

ing fault can be detected. Thus, we consider that a test subroutine is applied stimulating all monitored paths.

We implemented an AFS closed-loop with a threshold of 7 warnings: the clock frequency is increased when the number of warnings is less than 7 and it is decreased when there is more than 7 warnings. Higher this threshold, more energy efficient the circuit is. However, greater the risk of timing fault occurrence too. A regulation step of 1 MHz was adopted for the frequency generator.



(a) Complete AFS Simulink model



(b) Circuit Simulink model

Figure 4.8: (a) Simulink model for the complete AFS system, including the circuit itself and the AFS control. (b) Simulink model for the benchmark circuit, with the *Delay* and $\Delta p_{V_{th}}$ models for 80 circuit paths.

Figure 4.8 shows the Simulink model developed to simulate an AFS system. The first schema represents the complete AFS system while the the circuit block (the one at the middle) is depicted in the second schema. Both $\Delta p_{V_{th}}$ and *Delay* models are implemented inside the circuit block. Besides them, another block models the temperature in function of the supply voltage, the temperature and the activity. V and T measurements provided by a monitor are modelled as a Gaussian noise added to the real values of V and T, respectively. The monitor

flags are generated in the rightmost block by comparing the computed delays of the 80 paths with the clock period.

In the present implementation, we covered only the case where V is constant. We did not include voltage drops because their duration is usually in the order of nano- or microseconds, at the most. Here, a simulation step time of 1 millisecond has been adopted. The f_{MAX} at the beginning of the circuit lifetime is supposed equal to 3.1GHz for $V = 1.2V$ and 1.6GHz for $V = 0.8V$, at $T = 25\,^{\circ}C$. The *Delay* models can be observed through the occurrence of warnings. As it can be seen in Figure 4.9 for $V = 1.2V$, the frequency at which a warning is raised becomes lower for higher temperatures as well as for more degraded circuits.

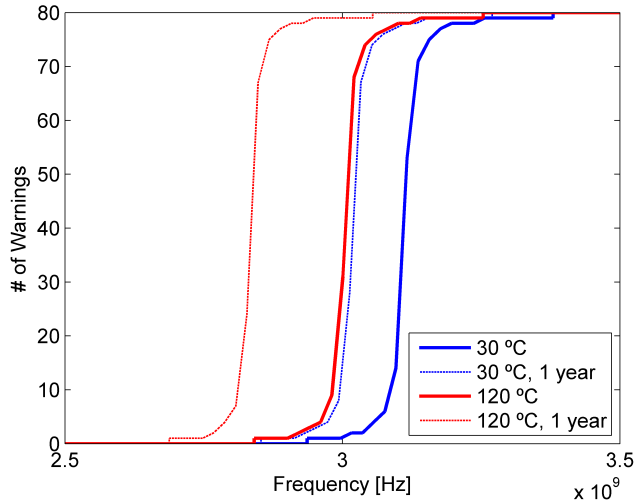


Figure 4.9: Number of warnings generated by the modelled delay monitors for $V = 1.2V$.

4.4.2 Aging degradation measurement for an AFS system

The proposed method for estimating aging is based on a conceptual AFS system with in-situ delay monitors and local temperature sensors. It consists basically in keeping record of both f_{MAX} and temperature evolution over time. The objective is to measure the shift of f_{MAX} exclusively due to aging, regardless of any voltage and temperature variations. In other words, it consists in extracting the $f_{MAX}(t)$ relationship from the $f_{MAX}(V, T, t)$ one, where V, T and t are the supply voltage, the temperature and the power-on time, respectively. Considering that we do not vary V in the present implementation, there is no need to find a $f_{MAX}(V)$ relationship. On the contrary, it is not possible to assume that each measure campaign would be conducted at or near the same temperature T since one does not have control over it. A $f_{MAX}(T)$ relationship must then be constructed before we can estimate $f_{MAX}(t)$.

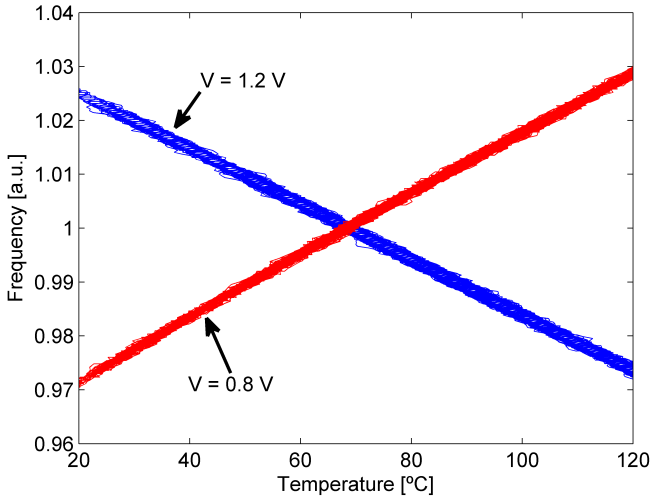


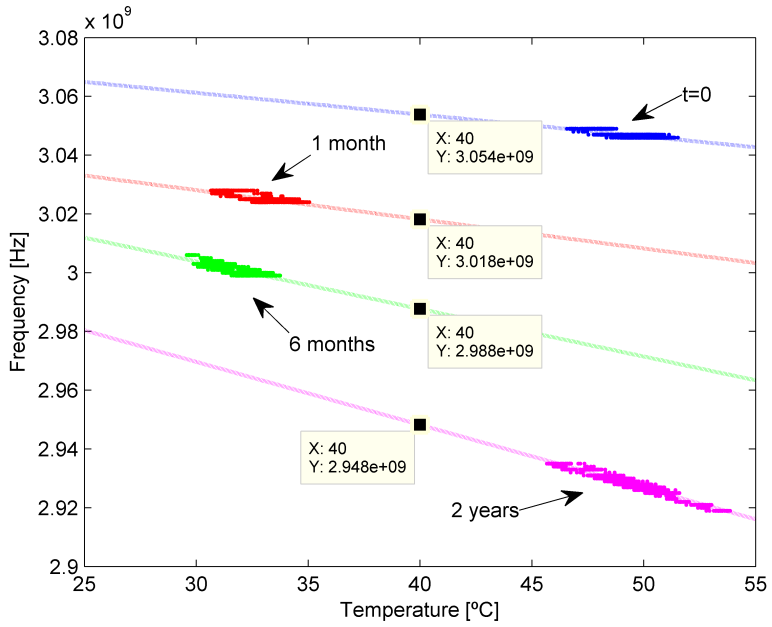
Figure 4.10: $f_{MAX}(T)$ relationship for $V = 1.2V$ (blue) and $V = 0.8V$ (red), normalized at $f_{MAX}(70\,^{\circ}C)$. The relationship is still linear but the slope sign changes due to the inverse temperature dependency.

A series of Simulink simulations with variable temperature was conducted. From the simulations, we found a linear $f_{MAX}(T)$ relationship for a wide temperature range, as shown in Figure 4.10 for both $V = 1.2V$ and $0.8V$. Even though its slope changes with V due to the inverse temperature dependency, the relationship is still linear. Therefore, $f_{MAX}(T)$ can be constructed with only 2 parameters:

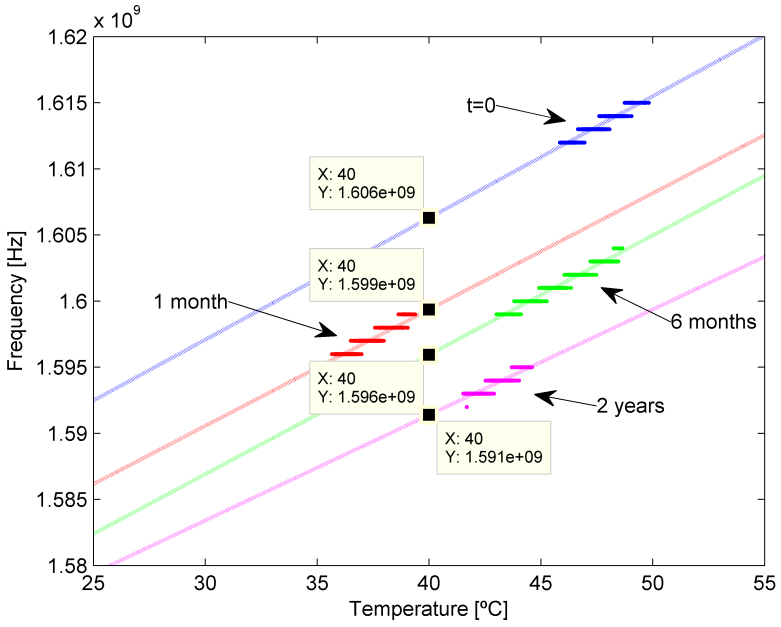
$$f_{MAX}(T) = a + b * T \tag{4.11}$$

Using equation(4.11), it is possible to calculate the aging-induced f_{MAX} shift even when the measure campaigns are conducted at different temperatures. A set of measures of both the frequency and the temperature is acquired each time an estimation of the degradation is needed. The linear $f_{MAX}(T)$ relationship is then constructed by applying linear regression on the obtained measures. In the end, the performance shift due to aging represented by $f_{MAX}(t)$ is obtained for any temperature using the estimations from previous measure campaigns.

Larger the number of measures done during a test campaign, more accurate the result. The average error of a $f_{MAX}(T)$ relationship constructed with 200 measures is 0.4% while it is less than 0.01% with 1000 points. The same analysis stands for the temperature range. $f_{MAX}(T)$ relationships constructed from measures made within a limited range of $2\,^{\circ}C$ and $3\,^{\circ}C$ presents an average residual of 1% and 0.1%, respectively. These small temperature ranges can be easily achieved by just varying the circuit activity, for example. In the following scenarios, we gathered 1000 measures within a temperature range of about $5\,^{\circ}C$. Each test campaign takes 1 second to be executed for a sample time of 1 ms.



(a) $V = 1.2V$



(b) $V = 0.8V$

Figure 4.11: Results of test campaigns conducted at (a) $V = 1.2V$ and (b) $V = 0.8V$ for 4 different stress times (1000 measures each). The dashed line is the associated $f_{MAX}(T)$. $f_{MAX}(40\,^{\circ}C)$ is reported for each stress time.

Figure 4.11 exemplifies the proposed solution through two different test case scenarios, the first one at $V = 1.2V$ and the second one at $V = 0.8V$. Each scenario reports the measures of 4 test campaigns performed at different times: beginning of lifetime (blue), after 1 month (red), 6 months (green) and 2 years (magenta). The dashed lines correspond to the associated $f_{MAX}(T)$. The tests were performed at different temperatures, from 30 $^\circ C$ to 50 $^\circ C$. Even so, we can estimate $f_{MAX}(t)$ at any temperature. For $V = 1.2V$, the aging-induced f_{MAX} shift is equal to -106 MHz (-3.47%) after 2 years at T = 40 $^\circ C$. For $V = 0.8V$, it is -15 MHz (0.93%). As expected, the circuit gets more degraded at a higher value of V. Note that, in the second scenario, f_{MAX} measured at t =6 months is actually higher than at t =1 month due to the temperature difference. It justifies the need of constructing a $f_{MAX}(T)$ relationship to find the aging-induced f_{MAX} shift.

Finally, the $f_{MAX}(t)$ relationship is gradually constructed during the circuit lifetime as more test campaigns are being performed. Though only 4 campaigns were performed in the examples presented here, in a real application they are to be conducted more often. As both BTI and HCI effects have an exponential dependence on the power-on time, a power function can be adopted for the $f_{MAX}(t)$ relationship:

$$f_{MAX}(t) = a - b * t^n \tag{4.12}$$

Coefficients a, b and n are computed by applying regression analysis on the data obtained from the test campaigns. For instance, Figure 4.12 shows the $f_{MAX}(t)$ relationships constructed considering the scenarios presented in Figure 4.11 for both $V = 1.2V$ and $V = 0.8V$ considering $T = 40°C$. The estimated coefficients of equation (4.12) are given in Table 4.2 for both $f_{MAX}(t)$ relationships. Through the constructed $f_{MAX}(t)$ relationship, it is then possible to predict f_{MAX} for any power-on time. In Figure 4.12, $f_{MAX}(t)$ has been predicted up to 5 years (60 months). Actually, the degradation rate will not necessarily remain constant since the operating conditions may change, namely, the supply voltage, temperature and workload. However, it provides an accurate estimate taking into account past operating conditions.

Table 4.2: Coefficients of the $f_{MAX}(t)$ relationship (equation (4.12)) computed for both scenarios presented in Figure 4.11. The resulting $f_{MAX}(t)$ curves are shown in Figure 4.12.

V	a	b	n
1.2V	3.05×10^9	2.35×10^5	0.34
0.8V	1.61×10^9	1.62×10^5	0.25

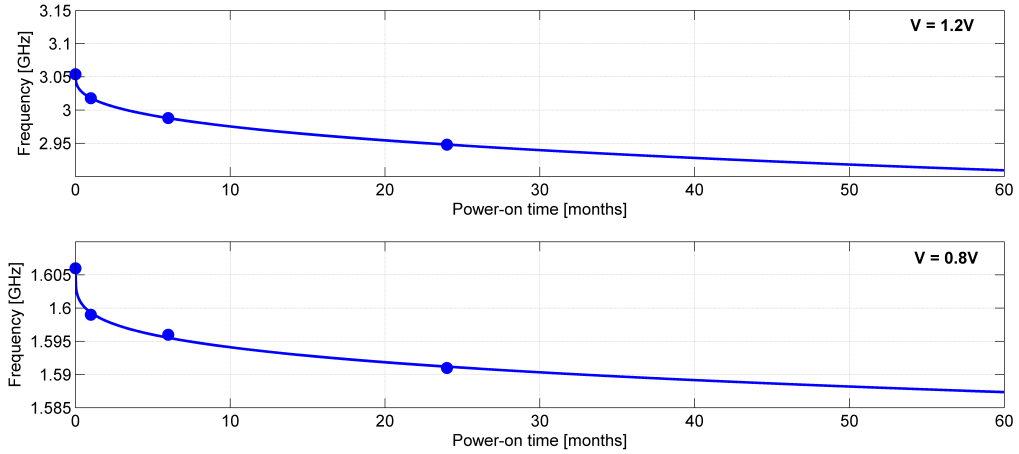


Figure 4.12: $f_{MAX}(t)$ relationship constructed with the data given in Figure 4.11 for $V = 1.2V$ (top) and $V = 0.8V$ (bottom), considering $T = 40°C$.

4.5 Conclusion

This chapter proposes and demonstrates some cases of application of the methodology presented in Chapter 3. First, the proposed models are used for on-line estimating the circuit degradation. They can be used in an adaptive system to track the maximum operating frequency. Since the computation of both models require some processor cycles, LUTs can be constructed in advance for achieving a fast response to variations. Otherwise, the circuit MTTF can be dynamically estimated taking into account the actual conditions of operation. Likewise, the maximum operating conditions can be also calculated on the fly to avoid a premature failure of the circuit. Lastly, the value of $\Delta p_{v_{th}}$ can serve as a reliability parameter to compare identical cores in a multi-core circuit to know which are the most degraded.

Besides of on-line estimating the circuit degradation, the proposed models can be used to perform reliability simulation of complex systems. Existing device-level aging models allow an accurate estimation of the aging-induced delay shift of a path. However, the simulation of several paths under different operating conditions (voltage, temperature, workload) is impossible due to the required simulation time. Therefore, this chapter demonstrated two applications of the proposed models for modelling a complex system.

The first application consisted in the implementation of a multi-core simulation framework in the MATLAB environment. This framework allows the evaluation of different task mapping strategies with regard to performance, energy and reliability. Here, five different strategies were implemented and compared. Through the simulation results, we could conclude that traditional DVFS strategies can lead to an increased aging degradation in spite of increasing both

performance and energy efficiency. As both BTI and HCI effects have an exponential dependence on the supply voltage, the circuit will endure less degradation if it operates at a medium voltage/frequency level than if it oscillates between low and high levels. Finally, it is not possible to have the best value possible in any of the indicators without worsening the other two. The best strategy is therefore the one that manages to reach a good balance between the three aspects, namely, performance, energy and reliability.

The second application was the modelling of a complete AFS system. Several critical paths of a benchmark circuit were modelled and their respective models were integrated in a Simulink model. Each model thus represented a path monitored by a slack time sensor in a conceptual AFS system. Then, a technique to estimate the circuit degradation of AFS systems for a given power-on time t was proposed. This technique consisted basically in keeping record of both f_{MAX} and temperature T variations over time. By creating a linear $f_{MAX}(T)$ relationship, it is then possible to calculate f_{MAX} for any temperature. It allows the estimation of the aging-induced performance shift even when the measures are not performed at a constant temperature. Finally, a $f_{MAX}(t)$ relationship is gradually constructed by fitting the obtained measures in a power function. The circuit f_{MAX} can then be predicted for any power-on time.

Chapter 5

Conclusion and Perspectives

5.1 Synthesis

The continuous miniaturization of transistor dimensions allows the design of compact circuits with even more processing capabilities and lower manufacturing costs. This has led to an exponential growth of the number of computing systems, in particular mobile devices. The energy consumed by these circuits must be minimized in order to extend their battery lifetime. Technology scaling has also considerably increased the circuit sensitivity to variations. Traditionally, the uncertainty in CMOS circuits is due to Process, Voltage and Temperature (PVT) variations, but aging effects have also become an important source of variability in recent technology nodes, particularly Bias Temperature Instability (BTI) and Hot Carrier Injection (HCI) mechanisms. The design of digital circuits nowadays requires the use of large voltage margins to ensure a functional circuit in the presence of variations. These margins, in turn, produce significant energy losses and underperformance.

Adaptive techniques have emerged in the last years to increase the energy efficiency of digital circuits by reducing the size of the safety guard-bands. Such techniques consist mainly in using in-situ sensors to monitor the dynamic variations of the maximum functional frequency f_{MAX}. Numerous f_{MAX} tracking techniques were already proposed in the literature, ranging from critical path replica to slack time monitors. However, all the existing solutions present the same limitation: they are not able to determine the source of variation.

While temperature and voltage variations cause temporary f_{MAX} shifts, aging degradation results in a permanent shift. Moreover, the use of dedicated sensors to estimate aging has not yet proved to be enough to provide an accurate information on the degradation of the logic circuit. Safety-critical systems, e.g. automotive and aerospace, require circuits that are able to operate over 25 years

in harsh environments with a very low failure rate. Handling aging effects in such systems poses then an important challenge.

Therefore, in this thesis we proposed a novel methodology to develop behavioral circuit-level aging models for digital logic circuits. Two models are issued. The first one, called *Delay*, estimates the critical path delay based on PVT variations. The second model, called $\Delta p_{V_{th}}$, reflects the impact of both BTI and HCI effects on the path delay. It is integrated into the *Delay* model as an aging-induced parameter shift. Built from SPICE simulations, these models take in account all factors that impact global aging, namely, circuit topology, voltage, temperature and workload. Since the effects of dynamic variations are considered, the models can be used for on-line estimation of the circuit performance. Finally, the methodology is architecture- and technology-independent, which means that it can be applied to any circuit while the obtained models depend on the circuit and the technology. After both the *Delay* and the $\Delta p_{V_{th}}$ formula are constructed for a given technology, the whole process (i.e. SPICE simulation and parameters identification) can be automated through scripts. This methodology has been described in Chapter 3.

An on-line estimation of the circuit maximum frequency f_{MAX} using the proposed models is only possible with sensors that provide reliable values of the voltage and the temperature. The main weakness of existing voltage and temperature monitors is that they are not robust to aging variations. Therefore, we considered here a small digital sensor composed of 7 ring-oscillators proposed in [100]. Its VT estimation method makes use of a database composed of the oscillating frequencies of all ring-oscillators for each $\{V, T\}$ condition. The frequencies are estimated at design phase. A calibration method is then applied after the circuit fabrication to update the database against process variations. However, the ring-oscillators are also sensitive to aging variations.

Therefore, we first evaluated the impact of BTI and HCI effects on the VT estimates. The aging effects result in a shift of the oscillating frequency of the ring-oscillators. In turn, this shift results in an offset on the voltage and temperature estimates. To counteract aging effects, we proposed a simple recalibration method. This method consists basically in remeasuring the frequencies of all ring-oscillators in a known $\{V, T\}$ condition. The measured frequency shifts (in Hertz) due to aging are then applied as a correction factor to all models in the database. We demonstrated that, by using the proposed recalibration method, the accuracy of the VT estimates for a 10 years old sensor (considering worst-case aging conditions) is almost equivalent to the one obtained with a fresh sensor. This contribution has been presented in Chapter 2.

Different scenarios of use have been considered for the proposed models. They have been shortly discussed in Chapter 4. For instance, they can be used in an adaptive control scheme to dynamically estimate f_{MAX}. To speed-up the

estimation process, a LUT can be previously constructed with the frequency shifts for small variations of the voltage and the temperature. Moreover, the models can be used to calculate online the circuit Mean Time to Failure (MTTF). The MTTF is usually computed at the design phase considering expected operating conditions (voltage, temperature and workload). Through the proposed models, it is then possible to estimate the MTTF on the fly considering actual operating conditions. Conversely, the maximum operating conditions for satisfying a given MTTF can be also dynamically estimated taking into account the current aging degradation. Finally, in a multi- or many-core context, the value of $\Delta p_{V_{th}}$ can be adopted as an aging indicator for reliability-aware task mapping strategies. By comparing the values of $\Delta p_{V_{th}}$, such strategies can favor the fresher cores over the more degraded ones when allocating the tasks.

In addition to on-line applications, the interest of the proposed models were also demonstrated through two off-line simulation contexts. The first consisted in a framework for simulating a multi-core circuit. This framework allows the evaluation of task mapping and DVFS strategies with regard to performance, energy and reliability. The performance is measured as the number of tasks completed before deadline, where the *Delay* model is used to estimate the required supply voltage for a given clock frequency, or *vice-versa*. Besides integrating the *Delay* model, $\Delta p_{V_{th}}$ is used as the reliability indicator. The energy is obtained from both dynamic and static power dissipation. At the end, five different task mapping and DVFS strategies were evaluated with the simulation framework. We could conclude that there is not a unique strategy that gives the best results in all the three parameters. The best strategy is the one that gives the most balanced trade-off between performance, energy and reliability.

The models were also incorporated in an AFS system simulated under Simulink. The Simulink model allows the simulation of AFS systems under different conditions of use. The simulated circuit contains slack monitors that warn the pre-occurrence of timing faults. The operation of these monitors is simulated through the use of the *Delay* model with the $\Delta p_{V_{th}}$ one to reproduce aging effects. The clock frequency of the circuit is increased or decreased based on the number of warning flags. Then, we proposed a technique to estimate the aging degradation of such systems. This technique consists in tracking the clock frequency and the temperature variations. By creating a linear relationship, the impact of temperature can be removed from the clock frequency fluctuation. An exponential relationship is finally constructed for the aging degradation using measurements of the clock frequency obtained over the circuit lifetime.

5.2 Perspectives

The results obtained in this thesis have confirmed the feasibility of modelling BTI and HCI effects at circuit-level from existing device-level models. As shown here, such simple but accurate aging models may have different applications, from on-line estimation of the circuit health to off-line evaluation of aging degradation of complex systems with variable operating conditions. However, there is still room for further improvements, as listed below.

Validation on silicon

The most important perspective is the validation of the proposed methodology on a real platform. For such, a circuit featuring f_{MAX} tracking sensors must be adopted in order to validate the accuracy of the proposed models. This would allow to compare the estimated propagation delay with the information provided by the embedded sensors. In addition to the model accuracy, the implementation costs in terms of computation, memory, area and power can be analyzed. Area and power overheads are mainly due to the required voltage and temperature sensors. Eventually, they may also be due to a co-processor designed to perform the computation of the models outputs. The memory overhead is due to the amount of memory space needed to store all the parameters for both the *Delay* and the Δp_{Vth} models. Finally, the computation overhead is defined by the amount of time that the processor is used to perform the computation of the models, if no dedicated hardware is used.

Near-threshold voltage modelling

Near-threshold voltage (NTV) design has emerged in the last years as a prominent solution for low-power circuits [150]. However, as stated in Chapter 1, the circuit is more sensitive to variations when operating close to the threshold voltage. The proposed models were validated here with supply voltages only down to 0.8V, while the threshold voltage in $28nm$ FD-SOI technology is smaller than 0.4V (depending on the temperature and body bias voltage) [151]. The Sakurai's alpha power law model [22], adopted as root for the *Delay* model, might be not valid for values of the supply voltage close to the threshold voltage. A validation of the proposed methodology with lower values of the supply voltage is therefore necessary if one wants to use it in NTV circuits.

Validation on different technologies

The proposed methodology has been validated on two different architectures, namely, a DSP and a RISC processor. Both circuits present different timing characteristics, the second one having critical paths about 3 times larger than the

first one. However, both circuits were designed in $28nm$ FD-SOI technology. Our methodology is assumed to be technology-independent, since CMOS circuits are all similar despite the transistor size. For instance, Sakurai proposed the alpha-power law model in 1990 [22] for the propagation delay of a CMOS inverter and we have demonstrated that it is still valid for current technologies. Nevertheless, the models might need to be redefined depending on the technology, but the overall methodology (SPICE simulation and data fitting) remains valid. The only limitation for applying our methodology to a different technology is that device-level models for both BTI and HCI effects must have been previously developed for the targeted technology.

Body bias voltage

In Section 2.2.1, we have highlighted the benefits in terms of energy efficiency provided by the use of Body Bias Voltage (Vbb) in adaptive techniques, especially in FD-SOI technologies. However, we have not explicitly incorporated it in the methodology proposed in this work. Despite the increased complexity, the integration of Vbb is straightforward. It consists only in an additional variable in both $Delay$ and $\Delta p_{V_{th}}$ models.

Workload dependence

The workload-dependence of $\Delta p_{V_{th}}$ could be further improved. In a complex processor it might be hard to know all the possible workloads during the design phase. This would make it difficult to obtain a set of $\Delta p_{V_{th}}$ for every possible workload. The best alternative seems to be the integration of the workload as a fourth variable (in addition to V, T and t). This fourth variable could be defined as the signal probability and/or activity of the source flip-flop of the modelled critical path. Otherwise, the workload could be determined from the combination of the signal probabilities of a set of representative flip-flops, as proposed in [144].

Random values of signal probabilities were adopted here to simulate different workloads. However, recent experiments performed in [145], with real benchmarks and input data, shown that the observed signal probabilities within the critical paths did not significantly vary between the benchmarks. As a consequence, the resulting NBTI-induced f_{MAX} shifts were almost identical. This means that NBTI is actually independent on the workload. As such assumption goes against the numerous previous work on NBTI dependence on the workload, e.g. [152, 136], it still requires further analysis before being adopted.

Bibliography

[1] Gordon Moore. Cramming more components onto integrated circuits. *Electronics*, 38(8), April 1965. [cited at p. ix]

[2] Ubergizmo. Samsung Galaxy S5 vs. Samsung Galaxy S8. [cited at p. ix]

[3] Intel. A Guide to the Internet of Things Infographic, October 2017. [cited at p. x]

[4] Semiconductor Industry Association (SIA) and Semiconductor Research Corporation (SRC). Rebooting the IT Revolution: A Call to Action. Technical report, September 2015. [cited at p. x]

[5] Ronald G. Dreslinski, Michael Wieckowski, David Blaauw, Dennis Sylvester, and Trevor Mudge. Near-Threshold Computing: Reclaiming Moore's Law Through Energy Efficient Integrated Circuits. *Proceedings of the IEEE*, 98(2):253–266, February 2010. [cited at p. x, 131]

[6] Didier Regis, Guillaume Hubert, Frank Bayle, and Marc Gatti. IC components reliability concerns for avionics end-users. In *Digital Avionics Systems Conference (DASC), 2013 IEEE/AIAA 32nd*, pages 2C2–1. IEEE, 2013. [cited at p. xi, 92]

[7] Mauricio Altieri, Suzanne Lesecq, Edith Beigne, and Olivier Heron. Towards online estimation of BTI/HCI-induced frequency degradation. In *Reliability Physics Symposium (IRPS), 2017 IEEE International*, 2017. [cited at p. xii, 46]

[8] Mauricio Altieri, Suzanne Lesecq, Edith Beign, and Olivier Heron. Method and device for estimating circuit aging, March 2017. [cited at p. xii, 46]

[9] Mauricio Altieri, Suzanne Lesecq, Diego Puschini, Olivier Heron, Edith Beigne, and Jorge Rodas. Evaluation and mitigation of aging effects on a digital on-chip voltage and temperature sensor. In *Power and Timing Modeling, Optimization and Simulation (PATMOS), 2015 25th International Workshop on*, pages 111–117. IEEE, 2015. [cited at p. xiii, 32]

[10] Mauricio Altieri, Suzanne Lesecq, Edith Beigne, Olivier Heron, and Diego Puschini. Tracking BTI and HCI effects at circuit-level in adaptive systems. In *New Circuits and Systems Conference (NEWCAS), 2016 14th IEEE International*, pages 1–4. IEEE, 2016. [cited at p. xiii, 86]

[11] Changhwan Shin. *Variation-Aware Advanced CMOS Devices and SRAM*, volume 56 of *Springer Series in Advanced Microelectronics*. Springer Netherlands, Dordrecht, 2016. DOI: 10.1007/978-94-017-7597-7. [cited at p. 2]

[12] Samar K. Saha. Modeling Process Variability in Scaled CMOS Technology. *IEEE Design & Test of Computers*, 27(2):8–16, March 2010. [cited at p. 2]

[13] Shekhar Borkar, Tanay Karnik, Siva Narendra, Jim Tschanz, Ali Keshavarzi, and Vivek De. Parameter variations and impact on circuits and microarchitecture. In *Proceedings of the 40th annual Design Automation Conference*, pages 338–342. ACM, 2003. [cited at p. 2, 3, 5, 7, 131]

[14] Martin Wirnshofer. *Variation-Aware Adaptive Voltage Scaling for Digital CMOS Circuits*, volume 41 of *Springer Series in Advanced Microelectronics*. Springer Netherlands, Dordrecht, 2013. DOI: 10.1007/978-94-007-6196-4. [cited at p. 3, 5]

[15] Drego Nigel Anthony. *Characterization and mitigation of process variation in digital circuits and systems*. PhD thesis, Massachusetts Institute of Technology, 2009. [cited at p. 4]

[16] Kelin Kuhn, Chris Kenyon, Avner Kornfeld, Mark Liu, Atul Maheshwari, Weikai Shih, Sam Sivakumar, Greg Taylor, Peter VanDerVoorn, and Keith Zawadzki. Managing Process Variation in Intel's 45nm CMOS Technology. *Intel Technology Journal*, 12(2), 2008. [cited at p. 4, 131]

[17] A. Asenov, S. Kaya, and A.R. Brown. Intrinsic parameter fluctuations in decananometer mosfets introduced by gate line edge roughness. *IEEE Transactions on Electron Devices*, 50(5):1254–1260, May 2003. [cited at p. 4]

[18] A. Asenov, S. Kaya, and J.H. Davies. Intrinsic threshold voltage fluctuations in decanano MOSFETs due to local oxide thickness variations. *IEEE Transactions on Electron Devices*, 49(1):112–119, January 2002. [cited at p. 4]

[19] M.J.M. Pelgrom, A.C.J. Duinmaijer, and A.P.G. Welbers. Matching properties of MOS transistors. *IEEE Journal of Solid-State Circuits*, 24(5):1433–1439, October 1989. [cited at p. 4]

[20] Yun Ye, Samatha Gummalla, Chi-Chao Wang, Chaitali Chakrabarti, and Yu Cao. Random variability modeling and its impact on scaled CMOS circuits. *Journal of Computational Electronics*, 9(3-4):108–113, December 2010. [cited at p. 4, 5, 131]

[21] K.L. Wong, T. Rahal-arabi, M. Ma, and G. Taylor. Enhancing Microprocessor Immunity to Power Supply Noise With Clock-Data Compensation. *IEEE Journal of Solid-State Circuits*, 41(4):749–758, April 2006. [cited at p. 6, 131]

[22] T. Sakurai and A.R. Newton. Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas. *IEEE Journal of Solid-State Circuits*, 25(2):584–594, April 1990. [cited at p. 7, 53, 82, 112, 113]

[23] Meeta S. Gupta, Jarod L. Oatley, Russ Joseph, Gu-Yeon Wei, and David M. Brooks. Understanding voltage variations in chip multiprocessors using a distributed power-delivery network. In *Design, Automation & Test in Europe Conference & Exhibition, 2007. DATE'07*, pages 1–6. IEEE, 2007. [cited at p. 7]

[24] Kerry Bernstein, David J. Frank, Anne E. Gattiker, Wilfried Haensch, Brian L. Ji, Sani R. Nassif, Edward J. Nowak, Dale J. Pearson, and Norman J. Rohrer. High-performance CMOS variability in the 65-nm regime and beyond. *IBM journal of research and development*, 50(4.5):433–449, 2006. [cited at p. 7]

[25] Hadi Esmaeilzadeh, Emily Blem, Renee St Amant, Karthikeyan Sankaralingam, and Doug Burger. Dark silicon and the end of multicore scaling. In *ACM SIGARCH Computer Architecture News*, volume 39, pages 365–376. ACM, 2011. [cited at p. 8, 94]

[26] Yoshio Miura and Yasuo Matukura. Investigation of Silicon-Silicon Dioxide Interface Using MOS Structure. *Japanese Journal of Applied Physics*, 5(2):180–180, February 1966. [cited at p. 9]

[27] Dieter K. Schroder and Jeff A. Babcock. Negative bias temperature instability: Road to cross in deep submicron silicon semiconductor manufacturing. *Journal of Applied Physics*, 94(1):1–18, July 2003. [cited at p. 9]

[28] V. Huard, M. Denais, and C. Parthasarathy. NBTI degradation: From physical mechanisms to modelling. *Microelectronics Reliability*, 46(1):1–23, January 2006. [cited at p. 9]

[29] Christian Schlunder, Stefano Aresu, Georg Georgakos, Werner Kanert, Hans Reisinger, Klaus Hofmann, and Wolfgang Gustin. HCI vs. BTI?-Neither one's out. In *Reliability Physics Symposium (IRPS), 2012 IEEE International*, pages 2F–4. IEEE, 2012. [cited at p. 9, 10, 32, 62]

[30] D.P. Ioannou, S. Mittl, and G. La Rosa. Positive Bias Temperature Instability Effects in nMOSFETs With $\hbox{HfO}_{2}/\hbox{TiN}$ Gate Stacks. *IEEE Transactions on Device and Materials Reliability*, 9(2):128–134, June 2009. [cited at p. 9]

[31] Vincent Huard. Two independent components modeling for negative bias temperature instability. In *Reliability Physics Symposium (IRPS), 2010 IEEE International*, pages 33–42. IEEE, 2010. [cited at p. 9, 65]

[32] Jyothi Bhaskarr Velamala, Ketul Sutaria, Takashi Sato, and Yu Cao. Physics matters: statistical aging prediction under trapping/detrapping. In *Proceedings of the 49th Annual Design Automation Conference*, pages 139–144. ACM, 2012. [cited at p. 9]

[33] Xinfei Guo, Wayne Burleson, and Mircea Stan. Modeling and Experimental Demonstration of Accelerated Self-Healing Techniques. In *Design Automation Conference (DAC), 2014 51st ACM/EDAC/IEEE*, pages 1–6. ACM Press, 2014. [cited at p. 9]

[34] F. Cacho, P. Mora, W. Arfaoui, X. Federspiel, and V. Huard. Hci/bti coupled model: The path for accurate and predictive reliability simulations. In *Reliability Physics Symposium (IRPS), 2014 IEEE International*, pages 5D–4. IEEE, 2014. [cited at p. 10, 37, 49, 52, 76, 82, 133]

[35] Xiaofei Wang, Qianying Tang, Pulkit Jain, Dong Jiao, and Chris H. Kim. The Dependence of BTI and HCI-Induced Frequency Degradation on Interconnect Length and Its Circuit Level Implications. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 23(2):280–291, February 2015. [cited at p. 10, 50, 62, 66, 67, 133]

[36] Stanislav Tyaginov, Markus Jech, Jacopo Franco, Prateek Sharma, Ben Kaczer, and Tibor Grasser. Understanding and Modeling the Temperature Behavior of Hot-Carrier Degradation in SiON nMOSFETs. *IEEE Electron Device Letters*, 37(1):84–87, January 2016. [cited at p. 10]

[37] Yao Wang, Sorin Cotofana, and Liang Fang. A unified aging model of NBTI and HCI degradation towards lifetime reliability management for nanoscale MOSFET circuits. In *Nanoscale Architectures (NANOARCH), 2011 IEEE/ACM International Symposium on*, pages 175–180. IEEE, June 2011. [cited at p. 10]

[38] Ben Kaczer, S. Mahato, V. Valduga de Almeida Camargo, M. Toledano-Luque, Ph J. Roussel, T. Grasser, Francky Catthoor, P. Dobrovolny, P. Zuber, G. Wirth, and others. Atomistic approach to variability of bias-temperature instability in circuit simulations. In *Reliability Physics Symposium (IRPS), 2011 IEEE International*, pages XT–3. Ieee, 2011. [cited at p. 10]

[39] D. Angot, Vincent Huard, L. Rahhal, A. Cros, Xavier Federspiel, A. Bajolet, Yann Carminati, M. Saliva, E. Pion, F. Cacho, and others. BTI variability fundamental understandings and impact on digital logic by the use of extensive dataset. In *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, pages 15–4, 2013. [cited at p. 10, 73, 74, 133]

[40] Christian Schlnder, Jrg Berthold, Fabian Proebster, Andreas Martin, Wolfgang Gustin, and Hans Reisinger. On the influence of BTI and HCI on parameter variability. In *Reliability Physics Symposium (IRPS), 2017 IEEE International*, pages 2E–4. IEEE, 2017. [cited at p. 10]

[41] Maria Toledano-Luque, Ben Kaczer, Jacopo Franco, Ph J. Roussel, T. Grasser, Thomas Y. Hoffmann, and Guido Groeseneken. From mean values to distributions of BTI lifetime of deeply scaled FETs through atomistic understanding of the degradation. In *VLSI Symp. Tech. Dig*, pages 152–153, 2011. [cited at p. 10, 11, 131]

[42] A. Kerber, P. Srinivasan, S. Cimino, P. Paliwoda, S. Chandrashekhar, Z. Chbili, S. Uppal, R. Ranjan, M.-I. Mahmud, D. Singh, and others. Device reliability metric for end-of-life performance optimization based on circuit level assessment. In *Reliability Physics Symposium (IRPS), 2017 IEEE International*, pages 2D–3. IEEE, 2017. [cited at p. 10]

[43] Richard Blish and Noel Durrant. Semiconductor Device Reliability Failure Models. In *Technology Transfer # 00053955A-XFR*. International Sematech, May 2000. [cited at p. 11]

[44] Edith Beigne, Pascal Vivet, Yvain Thonnart, Jean-Frederic Christmann, and Fabien Clermidy. Asynchronous Circuit Designs for the Internet of Everything: A Methodology for Ultralow-Power Circuits with GALS Architecture. *IEEE Solid-State Circuits Magazine*, 8(4):39–47, 2016. [cited at p. 14]

[45] John Logan. Statistical Circuit Design: Characterization and Modeling for Statistical Design. *Bell System Technical Journal*, 50(4):1105–1147, April 1971. [cited at p. 16]

[46] L. Stok and J. Koehl. Structured CAD: technology closure for modern ASICs [Tutorial]. In *Design, Automation & Test in Europe Conference & Exhibition, 2004. DATE'04*, pages xxxi–xxxi. IEEE Comput. Soc, 2004. [cited at p. 16]

[47] Robert. B. Hitchcock, Gordon L. Smith, and David D. Cheng. Timing Analysis of Computer Hardware. *IBM Journal of Research and Development*, 26(1):100–105, January 1982. [cited at p. 16]

[48] Cristiano Forzan and Davide Pandini. Why we need statistical static timing analysis. In *Computer Design, 2007. ICCD 2007. 25th International Conference on*, pages 91–96. IEEE, 2007. [cited at p. 16]

[49] Dirk P. Kroese, Tim Brereton, Thomas Taimre, and Zdravko I. Botev. Why the Monte Carlo method is so important today: Why the MCM is so important today. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(6):386–392, November 2014. [cited at p. 16]

[50] Cristiano Forzan and Davide Pandini. Statistical static timing analysis: A survey. *Integration, the VLSI Journal*, 42(3):409–435, June 2009. [cited at p. 16]

[51] V. Von Kaenel, P. Macken, and M.G.R. Degrauwe. A voltage reduction technique for battery-operated systems. *IEEE Journal of Solid-State Circuits*, 25(5):1136–1140, October 1990. [cited at p. 18, 22, 132]

[52] C. Niessen and B.C.H. Van. *An apparatus featuring a feedback signal for controlling a powering voltage for asynchronous electronic circuitry therein.* Google Patents, June 1993. [cited at p. 18]

[53] Lars Skovby Nielsen, Cees Niessen, Jens Sparso, and Kees Van Berkel. Low-power operation using self-timed circuits and adaptive scaling of the supply voltage. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2(4):391–397, 1994. [cited at p. 18]

[54] Tadahiro Kuroda, Kojiro Suzuki, Shinji Mita, Tetsuya Fujita, Fumiyuki Yamane, Fumihiko Sano, Akihiko Chiba, Yoshinori Watanabe, Koji Matsuda, Takeo Maeda, and others. Variable supply-voltage scheme for low-power high-speed CMOS digital design. *IEEE Journal of Solid-State Circuits*, 33(3):454–462, 1998. [cited at p. 18, 23, 132]

[55] Thomas D. Burd, Trevor A. Pering, Anthony J. Stratakos, and Robert W. Brodersen. A dynamic voltage scaled microprocessor system. *IEEE Journal of solid-state circuits*, 35(11):1571–1580, 2000. [cited at p. 18, 22, 132]

[56] B.H. Calhoun and A.P. Chandrakasan. Ultra-Dynamic Voltage Scaling (UDVS) Using Sub-Threshold Operation and Local Voltage Dithering. *IEEE Journal of Solid-State Circuits*, 41(1):238–245, January 2006. [cited at p. 19]

[57] E. Beigne, F. Clermidy, S. Miermont, A. Valentian, P. Vivet, S. Barasinski, F. Blisson, N. Kohli, and S. Kumar. A fully integrated power supply unit for fine grain power management application to embedded Low Voltage SRAMs. In *Solid-State Circuits Conference, 2008. ESSCIRC 2008. 34th European*, pages 138–141. IEEE, 2008. [cited at p. 19]

[58] Ivan Miro-Panades, Edith Beigne, Yvain Thonnart, Laurent Alacoque, Pascal Vivet, Suzanne Lesecq, Diego Puschini, Anca Molnos, Farhat Thabet, Benoit Tain, Karim Ben Chehida, Sylvain Engels, Robin Wilson, and Didier Fuin. A Fine-Grain Variation-Aware Dynamic Vdd-Hopping AVFS Architecture on a 32 nm GALS MPSoC. *IEEE Journal of Solid-State Circuits*, 49(7):1475–1486, July 2014. [cited at p. 19, 20, 36]

[59] James Tschanz, Nam Sung Kim, Saurabh Dighe, Jason Howard, Gregory Ruhl, Sriram Vangal, Siva Narendra, Yatin Hoskote, Howard Wilson, Carol Lam, Matthew Shuman, Carlos Tokunaga, Dinesh Somasekhar, Stephen Tang, David Finan, Tanay Karnik, Nitin Borkar, Nasser Kurd, and Vivek De. Adaptive Frequency and Biasing Techniques for Tolerance to Dynamic Temperature-Voltage Variations and Aging. In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2007 IEEE International*, pages 292–604. IEEE, February 2007. [cited at p. 19]

[60] James Tschanz, Keith Bowman, Shih-Lien Lu, Paolo Aseron, Muhammad Khellah, Arijit Raychowdhury, Bibiche Geuskens, Carlos Tokunaga, Chris Wilkerson, Tanay Karnik, and others. A 45nm resilient and adaptive microprocessor core for dynamic variation tolerance. In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2010 IEEE International*, pages 282–283. IEEE, 2010. [cited at p. 19]

[61] Edith Beigne, Alexandre Valentian, Ivan Miro-Panades, Robin Wilson, Philippe Flatresse, Fady Abouzeid, Thomas Benoist, Christian Bernard, Sebastien Bernard, Olivier Billoint, Sylvain Clerc, Bastien Giraud, Anuj Grover, Julien Le Coz, Jean-Philippe Noel, Olivier Thomas, and Yvain Thonnart. A 460 MHz at 397 mV, 2.6 GHz at 1.3 V, 32 bits VLIW DSP Embedding F MAX Tracking. *IEEE Journal of Solid-State Circuits*, 50(1):125–136, January 2015. [cited at p. 19, 21, 28, 50, 51, 82, 133, 134]

[62] R.B. Staszewski and P.T. Balsara. Phase-domain all-digital phase-locked loop. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 52(3):159–163, March 2005. [cited at p. 19, 20]

[63] Carolina Albea, Diego Puschini, Suzanne Lesecq, and Edith Beign. Optimal and robust control for a small-area FLL. In *Control & Automation (MED), 2011 19th Mediterranean Conference on*, pages 1100–1105. IEEE, 2011. [cited at p. 20]

[64] Davide Rossi, Antonio Pullini, Igor Loi, Michael Gautschi, Frank Kagan Gurkaynak, Adam Teman, Jeremy Constantin, Andreas Burg, Ivan Miro-Panades, Edith Beign, and others. 193 MOPS/mW@ 162 MOPS, 0.32 V to 1.15 V voltage range multi-core accelerator for energy efficient parallel and sequential digital processing. In *Low-Power and High-Speed Chips (COOL CHIPS XIX), 2016 IEEE Symposium in*, pages 1–3. Ieee, 2016. [cited at p. 20]

[65] Siva Narendra, Dimitri Antoniadis, and Vivek De. Impact of Using Adaptive Body Bias to Compensate Die-to-die Vt Variation on Within-die Vt variation. In *Low Power Electronics and Design (ISLPED), 1999 IEEE/ACM International Symposium on*, 1999. [cited at p. 20]

[66] Masayuki Miyazaki, Goichi Ono, Toshihiro Hattori, Kenji Shiozawa, Kunio Uchiyama, and Koichiro Ishibashi. A 1000-MIPS/W microprocessor using speed adaptive threshold-voltage CMOS with forward bias. In *Solid-State Circuits Conference, 2000. Digest of Technical Papers. ISSCC. 2000 IEEE International*, pages 420–421. IEEE, 2000. [cited at p. 20]

[67] J.W. Tschanz, J.T. Kao, S.G. Narendra, R. Nair, D.A. Antoniadis, A.P. Chandrakasan, and V. De. Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage. *IEEE Journal of Solid-State Circuits*, 37(11):1396–1402, November 2002. [cited at p. 20]

[68] Philippe Magarshack, Philippe Flatresse, and Giorgio Cesana. UTBB FD-SOI: A process/design symbiosis for breakthrough energy-efficiency. In *Proceedings of the Conference on Design, Automation and Test in Europe*, pages 952–957. EDA Consortium, 2013. [cited at p. 21]

[69] Davide Rossi, Antonio Pullini, Igor Loi, Michael Gautschi, Frank K. Grkaynak, Andrea Bartolini, Philippe Flatresse, and Luca Benini. A 60 GOPS/W, 1.8v to 0.9v body bias ULP cluster in 28nm UTBB FD-SOI technology. *Solid-State Electronics*, 117:170–184, March 2016. [cited at p. 21]

[70] Yeter Akgul, Diego Puschini, Suzanne Lesecq, Edith Beign, Ivan Miro-Panades, Pascal Benoit, and Lionel Torres. Power management through DVFS and dynamic body biasing in FD-SOI circuits. In *Proceedings of the 51st Annual Design Automation Conference*, pages 1–6. ACM, 2014. [cited at p. 21]

[71] Diego Puschini, Jorge Rodas, Edith Beigne, Mauricio Altieri, and Suzanne Lesecq. Body Bias usage in UTBB FDSOI designs: A parametric exploration approach. *Solid-State Electronics*, 117:138–145, March 2016. [cited at p. 21]

[72] X. Federspiel, D. Angot, M. Rafik, F. Cacho, A. Bajolet, N. Planes, D. Roy, M. Haond, and F. Arnaud. 28nm node bulk vs FDSOI reliability comparison. In *Reliability Physics Symposium (IRPS), 2012 IEEE International*, pages 3B–1. IEEE, 2012. [cited at p. 21]

[73] C. Ndiaye, V. Huard, X. Federspiel, F. Cacho, and A. Bravaix. Performance vs. reliability adaptive body bias scheme in 28nm & 14nm UTBB FDSOI nodes. *Microelectronics Reliability*, 64:158–162, September 2016. [cited at p. 21]

[74] Milovan Blagojevi, Martin Cochet, Ben Keller, Philippe Flatresse, Andrei Vladimirescu, and Borivoje Nikoli. A fast, flexible, positive and negative adaptive body-bias generator in 28nm FDSOI. In *VLSI Circuits (VLSI-Circuits), 2016 IEEE Symposium on*, pages 1–2. IEEE, 2016. [cited at p. 21]

[75] Masato Nagamatsu, H. Tago, T. Mijamori, M. Kamata, H. Murakami, Y. Ootaguro, H. Goto, T. Utsumi, T. Teruyama, K. Mabuchi, and others. A 150 MIPS/W CMOS RISC processor for PDA applications. In *Solid-State Circuits*

Conference, 1995. Digest of Technical Papers. 41st ISSCC, 1995 IEEE International, pages 114–115. IEEE, 1995. [cited at p. 23]

[76] Dan Ernst, Nam Sung Kim, Shidhartha Das, Sanjay Pant, Rajeev Rao, Toan Pham, Conrad Ziesler, David Blaauw, Todd Austin, Krisztian Flautner, and others. Razor: A low-power pipeline based on circuit-level timing speculation. In *Microarchitecture, 2003. MICRO-36. Proceedings. 36th Annual IEEE/ACM International Symposium on*, pages 7–18. IEEE, 2003. [cited at p. 24]

[77] Shidhartha Das, Carlos Tokunaga, Sanjay Pant, Wei-Hsiang Ma, Sudherssen Kalaiselvan, Kevin Lai, David M. Bull, and David T. Blaauw. RazorII: In Situ Error Detection and Correction for PVT and SER Tolerance. *IEEE Journal of Solid-State Circuits*, 44(1):32–48, January 2009. [cited at p. 24, 132]

[78] P.E. Dodd and L.W. Massengill. Basic mechanisms and modeling of single-event upset in digital microelectronics. *IEEE Transactions on Nuclear Science*, 50(3):583–602, June 2003. [cited at p. 24]

[79] Matthew Fojtik, David Fick, Yejoong Kim, Nathaniel Pinckney, David Money Harris, David Blaauw, and Dennis Sylvester. Bubble Razor: Eliminating Timing Margins in an ARM Cortex-M3 Processor in 45 nm CMOS Using Architecturally Independent Error Detection and Correction. *IEEE Journal of Solid-State Circuits*, 48(1):66–81, January 2013. [cited at p. 25, 132]

[80] Shuam Sadasivan. An Introduction to the ARM Cortex-M3 Processor. *ARM Holdings*, October 2006. [cited at p. 25]

[81] Mridul Agarwal, Bipul C. Paul, Ming Zhang, and Subhasish Mitra. Circuit failure prediction and its application to transistor aging. In *VLSI Test Symposium, 2007. 25th IEEE*, pages 277–286. IEEE, 2007. [cited at p. 26, 27]

[82] Liangzhen Lai, Vikas Chandra, Robert Aitken, and Puneet Gupta. Slackprobe: A low overhead in situ on-line timing slack monitoring methodology. In *Proceedings of the Conference on Design, Automation and Test in Europe*, pages 282–287. EDA Consortium, 2013. [cited at p. 26, 27]

[83] Vincent Huard, F. Cacho, F. Giner, M. Saliva, A. Benhassain, Dinesh Patel, N. Torres, S. Naudet, Abhishek Jain, and C. Parthasarathy. Adaptive Wearout Management with in-situ aging monitors. In *Reliability Physics Symposium (IRPS), 2014 IEEE International*, pages 6B–4. IEEE, 2014. [cited at p. 26, 27, 85, 88, 101, 132]

[84] Ivan Miro-Panades, Edith Beigne, Olivier Billoint, and Yvain Thonnart. In-situ Fmax/Vmin tracking for energy efficiency and reliability optimization. In *On-Line Testing and Robust System Design (IOLTS), 2017 IEEE 23rd International Symposium on*, July 2017. [cited at p. 27, 28, 132]

[85] Eren Kursun and Chen-Yong Cher. Variation-aware thermal characterization and management of multi-core architectures. In *Computer Design, 2008. ICCD 2008. IEEE International Conference on*, pages 280–285. IEEE, 2008. [cited at p. 29]

[86] K. K. Kim, F. Ge, and K. Choi. On-chip process variation monitoring circuit based on gate leakage sensing. *Electronics letters*, 46(3):227–228, 2010. [cited at p. 29]

[87] Islam A. K. M. Mahfuzul, Akira Tsuchiya, Kazutoshi Kobayashi, and Hidetoshi Onodera. Variation-Sensitive Monitor Circuits for Estimation of Global Process Parameter Variation. *IEEE Transactions on Semiconductor Manufacturing*, 25(4):571–580, November 2012. [cited at p. 29]

[88] Young-Jae An, Dong-Hoon Jung, Kyungho Ryu, Hyuck Sang Yim, and Seong-Ook Jung. All-Digital ON-Chip Process Sensor Using Ratioed Inverter-Based Ring Oscillator. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, pages 1–11, 2016. [cited at p. 29]

[89] Tseng-Chin Luo, Mango C.-T. Chao, Michael S.-Y. Wu, Kuo-Tsai Li, Chin. C. Hsia, Huan-Chi Tseng, Chuen-Uan Huang, Yuan-Yao Chang, Samuel C. Pan, and Konrad K.-L. Young. A novel array-based test methodology for local process variation monitoring. In *Test Conference (ITC), 2009. International*, pages 1–9. IEEE, November 2009. [cited at p. 29]

[90] Kamran Souri, Youngcheol Chae, and Kofi A. A. Makinwa. A CMOS Temperature Sensor With a Voltage-Calibrated Inaccuracy of $\pm$ 0.15$ ˆ{\circ}$ C (3σ) From $-$ 55$ˆ{\circ}$ C to 125$ˆ{\circ}$ C. *IEEE Journal of Solid-State Circuits*, 48(1):292–301, January 2013. [cited at p. 29]

[91] A. Muhtaroglu, G. Taylor, and T. Rahal-Arabi. On-Die Droop Detector for Analog Sensing of Power Supply Noise. *IEEE Journal of Solid-State Circuits*, 39(4):651–660, April 2004. [cited at p. 29]

[92] Rex Petersen, Pankaj Pant, Pablo Lopez, Aaron Barton, Jim Ignowski, and Doug Josephson. Voltage transient detection and induction for debug and test. In *Test Conference, 2009. ITC 2009. International*, pages 1–10. IEEE, 2009. [cited at p. 29]

[93] I.M. Filanovsky and Su Tam Lim. Temperature sensor applications of diode-connected MOS transistors. In *Circuits and Systems (ISCAS), 2002 IEEE International Symposium on*, pages II–149–II–152. IEEE, 2002. [cited at p. 29]

[94] Poki Chen, Chun-Chi Chen, Chin-Chung Tsai, and Wen-Fu Lu. A time-to-digital-converter-based CMOS smart temperature sensor. *IEEE Journal of Solid-State Circuits*, 40(8):1642–1648, August 2005. [cited at p. 29]

[95] Eduardo Boemo and Sergio Lpez-Buedo. Thermal monitoring on FPGAs using ring-oscillators. In *Field-Programmable Logic and Applications*, pages 69–78. Springer, 1997. [cited at p. 29]

[96] S. Lopez-Buedo, J. Garrido, and E. Boemo. Thermal testing on reconfigurable computers. *IEEE Design & Test of Computers*, 17(1):84–91, March 2000. [cited at p. 29]

[97] Rajarshi Mukherjee, Somsubhra Mondal, and Seda Ogrenci Memik. Thermal sensor allocation and placement for reconfigurable systems. In *Computer-Aided Design, 2006. ICCAD'06. IEEE/ACM International Conference on*, pages 437–442. IEEE, 2006. [cited at p. 29]

[98] John J. Len Franco, Eduardo Boemo, Encarnacin Castillo, and Luis Parrilla. Ring oscillators as thermal sensors in FPGAs: Experiments in low voltage. In *Programmable Logic Conference (SPL), 2010 VI Southern*, pages 133–137. IEEE, 2010. [cited at p. 29]

[99] Basab Datta and Wayne Burleson. Low-power and robust on-chip thermal sensing using differential ring oscillators. In *Circuits and Systems, 2007. MWSCAS 2007. 50th Midwest Symposium on*, pages 29–32. IEEE, 2007. [cited at p. 30]

[100] Lionel Vincent, Philippe Maurine, Suzanne Lesecq, and Edith Beign. Embedding statistical tests for on-chip dynamic voltage and temperature monitoring. In *Proceedings of the 49th Annual Design Automation Conference*, pages 994–999. ACM, 2012. [cited at p. 30, 32, 35, 36, 43, 110]

[101] Yousuke Miyake, Yasuo Sato, Seiji Kajihara, and Yukiya Miura. Temperature and Voltage Measurement for Field Test Using an Aging-Tolerant Monitor. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 24(11):3282–3295, November 2016. [cited at p. 30, 32]

[102] John Keane, Xiaofei Wang, Devin Persaud, and Chris H. Kim. An All-In-One Silicon Odometer for Separately Monitoring HCI, BTI, and TDDB. *IEEE Journal of Solid-State Circuits*, 45(4):817–829, April 2010. [cited at p. 30, 31, 132]

[103] P. Singh, E. Karl, D. Blaauw, and D. Sylvester. Compact Degradation Sensors for Monitoring NBTI and Oxide Degradation. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 20(9):1645–1655, September 2012. [cited at p. 31]

[104] John Keane, Shrinivas Venkatraman, Paulo Butzen, and Chris H. Kim. An array-based test circuit for fully automated gate dielectric breakdown characterization. In *Custom Integrated Circuits Conference (CICC), 2008 IEEE*, pages 121–124. IEEE, September 2008. [cited at p. 31]

[105] Karl Hofmann, Hans Reisinger, K. Ermisch, C. Schlnder, Wolfgang Gustin, T. Pompl, Georg Georgakos, K. v Arnim, J. Hatsch, T. Kodytek, and others. Highly accurate product-level aging monitoring in 40nm CMOS. In *VLSI Technology (VLSIT), 2010 Symposium on*, pages 27–28. IEEE, 2010. [cited at p. 31]

[106] Jae-Joon Kim, Rahul M. Rao, Jeremy Schaub, Amlan Ghosh, Aditya Bansal, Kai Zhao, Barry P. Linder, and James Stathis. PBTI/NBTI monitoring ring oscillator circuits with on-chip Vt characterization and high frequency AC stress capability. In *VLSI Circuits (VLSIC), 2011 Symposium on*, pages 224–225. IEEE, 2011. [cited at p. 31]

[107] Hiromitsu Awano, Masayuki Hiromoto, and Takashi Sato. BTIarray: A Time-Overlapping Transistor Array for Efficient Statistical Characterization of Bias Temperature Instability. *IEEE Transactions on Device and Materials Reliability*, 14(3):833–843, September 2014. [cited at p. 31]

[108] Xiaoxiao Wang, LeRoy Winemberg, Donglin Su, Dat Tran, Saji George, Nisar Ahmed, Steve Palosh, Allan Dobin, and Mohammad Tehranipoor. Aging Adaption in Integrated Circuits Using a Novel Built-In Sensor. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 34(1):109–121, January 2015. [cited at p. 31]

[109] Tony Tae-Hyoung Kim, Pong-Fei Lu, Keith A. Jenkins, and Chris H. Kim. A Ring-Oscillator-Based Reliability Monitor for Isolated Measurement of NBTI and

PBTI in High-k/Metal Gate Technology. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 23(7):1360–1364, July 2015. [cited at p. 31, 85, 88]

[110] Yong Zhao and Hans G. Kerkhoff. Highly Dependable Multi-processor SoCs Employing Lifetime Prediction Based on Health Monitors. In *Test Symposium (ATS), 2016 25th Asian*, pages 228–233. IEEE, November 2016. [cited at p. 31]

[111] Deepashree Sengupta and Sachin Sapatnekar. Estimating Circuit Aging due to BTI and HCI using Ring-Oscillator-Based Sensors. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, pages 1–1, 2017. [cited at p. 31, 85, 88]

[112] V. Huard, E. Pion, F. Cacho, D. Croain, V. Robert, R. Delater, P. Mergault, S. Engels, P. Flatresse, N. Ruiz Amador, and others. A predictive bottom-up hierarchical approach to digital system reliability. In *Reliability Physics Symposium (IRPS), 2012 IEEE International*, pages 4B–1. IEEE, 2012. [cited at p. 31, 32, 37, 132]

[113] Lionel Vincent, Edith Beigne, Laurent Alacoque, Suzanne Lesecq, Catherine Bour, and Philippe Maurine. A fully integrated 32 nm multiprobe for dynamic PVT measurements within complex digital SoC. In *VARI: International Workshop on CMOS Variability*, 2011. [cited at p. 32]

[114] Lionel Vincent, Edith Beigne, Suzanne Lesecq, Julien Mottin, David Coriat, and Philippe Maurine. Dynamic Variability Monitoring Using Statistical Tests for Energy Efficient Adaptive Architectures. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 61(6):1741–1754, June 2014. [cited at p. 32, 33, 132]

[115] Lionel Vincent, Philippe Maurine, Edith Beign, Suzanne Lesecq, and Julien Mottin. Temperature and fast voltage on-chip monitoring using low-cost digital sensors. In *VARI: Workshop on CMOS Variability*, 2013. [cited at p. 34, 35, 36]

[116] Mohamed Selim, Eric Jeandeau, and Cyril Descleves. DesignReliability Flow and Advanced Models Address IC-Reliability Issues. In *Early Reliability Modeling for Aging and Variability in Silicon Systems (ERMAVSS), Workshop on*, 2016. [cited at p. 37, 51]

[117] Shuo Wang, Jifeng Chen, and Mohammad Tehranipoor. Representative critical reliability paths for low-cost and accurate on-chip aging evaluation. In *Proceedings of the International Conference on Computer-Aided Design*, pages 736–741. ACM, 2012. [cited at p. 48]

[118] Chiara Sandionigi and Olivier Heron. Identifying aging-aware representative paths in processors. In *On-Line Testing Symposium (IOLTS), 2015 IEEE 21st International*, pages 32–33. IEEE, 2015. [cited at p. 48]

[119] Laurence W. Nagel. *SPICE2: A Computer Program to Simulate Semiconductor Circuits*. PhD thesis, EECS Department, University of California, Berkeley, 1975. [cited at p. 48]

[120] MATLAB. *version 8.3.0 (R2014a)*. The MathWorks Inc., Natick, MA, USA, February 2014. [cited at p. 49]

[121] X. Federspiel, F. Cacho, and D. Roy. Experimental characterization of the interactions between HCI, off-state and BTI degradation modes. In *Integrated Reliability Workshop Final Report (IRW), 2011 IEEE International*, pages 133–136. IEEE, 2011. [cited at p. 49]

[122] W. Arfaoui, X. Federspiel, P. Mora, M. Rafik, D. Roy, and A. Bravaix. Experimental analysis of defect nature and localization under hot-carrier and bias temperature damage in advanced CMOS nodes. In *Integrated Reliability Workshop Final Report (IRW), 2013 IEEE International*, pages 78–83. IEEE, 2013. [cited at p. 49]

[123] S. Mahapatra, V. Huard, A. Kerber, V. Reddy, S. Kalpat, and A. Haggag. Universality of NBTI-From devices to circuits and products. In *Reliability Physics Symposium (IRPS), 2014 IEEE International*, pages 3B–1. IEEE, 2014. [cited at p. 49, 50, 62, 63, 76]

[124] Veit B. Kleeberger, Martin Barke, Christoph Werner, Doris Schmitt-Landsiedel, and Ulf Schlichtmann. A compact model for NBTI degradation and recovery under use-profile variations and its application to aging analysis of digital integrated circuits. *Microelectronics Reliability*, 54(6-7):1083–1089, June 2014. [cited at p. 50, 62, 64, 75, 77, 78, 83, 134]

[125] Chenyue Ma, Hans Jurgen Mattausch, Kazuya Matsuzawa, Seiichiro Yamaguchi, Teruhiko Hoshida, Masahiro Imade, Risho Koh, Takahiko Arakawa, and Mitiko Miura-Mattausch. Universal NBTI Compact Model for Circuit Aging Simulation under Any Stress Conditions. *IEEE Transactions on Device and Materials Reliability*, 14(3):818–825, September 2014. [cited at p. 50, 62, 64]

[126] Keitul Sutaria, Jyothi Velamala, Chris H. Kim, Takashi Sato, and Yu Cao. Aging Statistics Based on Trapping/Detrapping: Compact Modeling and Silicon Validation. *IEEE Transactions on Device and Materials Reliability*, 14(2):607–615, June 2014. [cited at p. 50, 62]

[127] Mentor Graphics. Questa Advanced Simulator, July 2017. [cited at p. 50]

[128] Ali Dasdan and Ivan Hom. Handling inverted temperature dependence in static timing analysis. *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, 11(2):306–324, 2006. [cited at p. 53]

[129] Mitsuhiko Igarashi, Kan Takeuchi, Takeshi Okagaki, Koji Shibutani, Hiroaki Matsushita, and Koji Nii. An on-die digital aging monitor against HCI and xBTI in 16 nm Fin-FET bulk CMOS technology. In *European Solid-State Circuits Conference (ESSCIRC), ESSCIRC 2015-41st*, pages 112–115. IEEE, 2015. [cited at p. 62]

[130] Chen Zhou, Xiaofei Wang, Weichao Xu, Yuhao Zhu, Vijay Janapa Reddi, and Chris H. Kim. Estimation of instantaneous frequency fluctuation in a fast DVFS environment using an empirical BTI stress-relaxation model. In *Reliability Physics Symposium (IRPS), 2014 IEEE International*, pages 2D–2. IEEE, 2014. [cited at p. 64]

[131] S. Mahapatra, N. Goel, S. Desai, S. Gupta, B. Jose, S. Mukhopadhyay, K. Joshi, A. Jain, A. E. Islam, and M. A. Alam. A Comparative Study of Different Physics-Based NBTI Models. *IEEE Transactions on Electron Devices*, 60(3):901–916, March 2013. [cited at p. 65]

[132] Muhammad A. Alam. A critical examination of the mechanics of dynamic NBTI for PMOSFETs. In *Electron Devices Meeting, 2003. IEDM'03 Technical Digest. IEEE International*, pages 14–4. IEEE, 2003. [cited at p. 66]

[133] Ben Kaczer, Tibor Grasser, Ph J. Roussel, Jacopo Franco, Robin Degraeve, L.-A. Ragnarsson, Eddy Simoen, Guido Groeseneken, and Hans Reisinger. Origin of NBTI variability in deeply scaled pFETs. In *Reliability Physics Symposium (IRPS), 2010 IEEE International*, pages 26–32. IEEE, 2010. [cited at p. 66]

[134] Charly Bechara, Aurelien Berhault, Nicolas Ventroux, Stphane Chevobbe, Yves Lhuillier, Raphal David, and Daniel Etiemble. A small footprint interleaved multithreaded processor for embedded systems. In *Electronics, Circuits and Systems (ICECS), 2011 18th IEEE International Conference on*, pages 685–690. IEEE, 2011. [cited at p. 81, 82, 83, 90, 96, 134, 137]

[135] Jianxin Fang and Sachin S. Sapatnekar. The impact of BTI variations on timing in digital logic circuits. *IEEE Transactions on Device and Materials Reliability*, 13(1):277–286, 2013. [cited at p. 85]

[136] Ajith Sivadasan, S. Mhira, Armelle Notin, A. Benhassain, V. Huard, Etienne Maurin, F. Cacho, L. Anghel, and A. Bravaix. Architecture-and workload-dependent digital failure rate. In *Reliability Physics Symposium (IRPS), 2017 IEEE International*, pages CR–8. IEEE, 2017. [cited at p. 85, 113]

[137] V. Huard, S. Mhira, M. De Tomasi, E. Trabace, R. Enrici Vaion, and P. Zabberoni. Robust automotive products in advanced CMOS nodes. In *Reliability Physics Symposium (IRPS), 2017 IEEE International*, pages 3A–2. IEEE, 2017. [cited at p. 85]

[138] B. Rebaud, M. Belleville, E. Beign, C. Bernard, M. Robert, P. Maurine, and N. Azemard. Timing slack monitoring under process and environmental variations: Application to a DSP performance optimization. *Microelectronics Journal*, 42(5):718–732, May 2011. [cited at p. 85, 88, 101]

[139] Liangzhen Lai, Vikas Chandra, Robert C. Aitken, and Puneet Gupta. SlackProbe: A Flexible and Efficient In Situ Timing Slack Monitoring Methodology. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 33(8):1168–1179, August 2014. [cited at p. 85, 88, 101]

[140] Xiaofei Wang, John Keane, Tony Tae-Hyoung Kim, Pulkit Jain, Qianying Tang, and Chris H. Kim. Silicon odometers: Compact in situ aging sensors for robust system design. *IEEE Micro*, 34(6):74–85, 2014. [cited at p. 85, 88]

[141] Sergei Kostin, Jaan Raik, Raimund Ubar, Maksim Jenihhin, Thiago Copetti, Fabian Vargas, and Leticia Bolzani Poehls. SPICE-Inspired Fast Gate-Level Computation of NBTI-induced Delays in Nanoscale Logic. In *Design and Diagnostics of Electronic Circuits & Systems (DDECS), 2015 IEEE 18th International Symposium on*, pages 223–228. IEEE, April 2015. [cited at p. 87]

[142] Dominik Lorenz, Georg Georgakos, and Ulf Schlichtmann. Aging analysis of circuit timing considering NBTI and HCI. In *On-Line Testing Symposium, 2009. IOLTS 2009. 15th IEEE International*, pages 3–8. IEEE, 2009. [cited at p. 87]

[143] Zheng Wang, Shazia Kanwal, Lai Wang, and Anupam Chattopadhyay. Automated High-level Modeling of Power, Temperature and Timing Variation for Microprocessor. *The Journal of King Mongkut's University of Technology North Bangkok*, August 2017. [cited at p. 87]

[144] Arunkumar Vijayan, Abhishek Koneru, Saman Kiamehr, Krishnendu Chakrabarty, and Mehdi B. Tahoori. Fine-Grained Aging-Induced Delay Prediction Based on the Monitoring of Run-Time Stress. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, pages 1–1, 2016. [cited at p. 87, 113]

[145] Yukai Chen, Enrico Macii, and Massimo Poncino. Empirical derivation of upper and lower bounds of NBTI aging for embedded cores. *Microelectronics Reliability*, July 2017. [cited at p. 87, 113]

[146] Anup Das, Akash Kumar, and Bharadwaj Veeravalli. Reliability and Energy-Aware Mapping and Scheduling of Multimedia Applications on Multiprocessor Systems. *IEEE Transactions on Parallel and Distributed Systems*, 27(3):869–884, March 2016. [cited at p. 95]

[147] Jacopo Panerati and Giovanni Beltrame. Trading off power and fault-tolerance in real-time embedded systems. In *Adaptive Hardware and Systems (AHS), 2015 NASA/ESA Conference on*, pages 1–8. IEEE, 2015. [cited at p. 95]

[148] Haeseung Lee, Muhammad Shafique, and Mohammad Abdullah Al Faruque. Low-overhead Aging-aware Resource Management on Embedded GPUs. In *Design Automation Conference (DAC), 2017 54th ACM/EDAC/IEEE*, 2017. [cited at p. 95]

[149] A. Benhassain, F. Cacho, V. Huard, S. Mhira, L. Anghel, C. Parthasarathy, A. Jain, and A. Sivadasan. Robustness of timing in-situ monitors for AVS management. In *Reliability Physics Symposium (IRPS), 2016 IEEE International*, pages CR–4. IEEE, 2016. [cited at p. 101]

[150] Himanshu Kaul, Mark Anders, Steven Hsu, Amit Agarwal, Ram Krishnamurthy, and Shekhar Borkar. Near-threshold voltage (NTV) designOpportunities and challenges. In *Design Automation Conference (DAC), 2012 49th ACM/EDAC/IEEE*, pages 1149–1154. IEEE, 2012. [cited at p. 112]

[151] David Jacquet, Frederic Hasbani, Philippe Flatresse, Robin Wilson, Franck Arnaud, Giorgio Cesana, Thierry Di Gilio, Christophe Lecocq, Tanmoy Roy, Amit Chhabra, Chiranjeev Grover, Olivier Minez, Jacky Uginet, Guy Durieu, Cyril Adobati, Davide Casalotto, Frederic Nyer, Patrick Menut, Andreia Cathelin, Indavong Vongsavady, and Philippe Magarshack. A 3 GHz Dual Core Processor ARM Cortex TM -A9 in 28 nm UTBB FD-SOI CMOS With Ultra-Wide Voltage Range and Energy Efficiency Optimization. *IEEE Journal of Solid-State Circuits*, 49(4):812–826, April 2014. [cited at p. 112]

[152] Olivier Hron, Chiara Sandionigi, E. Piriou, S. Mbarek, and V. Huard. Workload-dependent BTI analysis in a processor core at high level. In *Reliability Physics Symposium (IRPS), 2015 IEEE International*, pages CA–6. IEEE, 2015. [cited at p. 113]

List of Publications

International Journals

1. D. Puschini, J. Rodas, E. Beigne, M. Altieri, S. Lesecq. Body Bias usage in UTBB FDSOI designs: A parametric exploration approach. In *Solid-State Electronics*, vol. 117, pp. 138-145, 2016.

International Conferences

1. C. Sandionigi, M. Altieri, O. Heron. Early estimation of aging in the design flow of integrated circuits through a programmable hardware module. In *DFT'17: IEEE Int. Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems*, Cambridge, UK, 2017. (In press).

2. M. Altieri, S. Lesecq, E. Beigne, O. Heron. Towards on-line estimation of aging in digital circuits through circuit-level models. In *IRPS'17: IEEE International Reliability Physics Symposium*, Monterey, USA, 2017.

3. M. Altieri, S. Lesecq, E. Beigne, O. Heron, D. Puschini. Tracking BTI and HCI effects at circuit-level in adaptive systems. In *NEWCAS'16: 14th IEEE International NEW Circuits And Systems conference*, Vancouver, Canada, 2016.

4. M. Altieri, S. Lesecq, D. Puschini, O. Heron, E. Beigne. Evaluation and mitigation of aging effects on a digital on-chip voltage and temperature sensor. In *PATMOS'15: 25th International Workshop on Power and Timing Modeling, Optimization and Simulation*, Salvador, Brazil, 2015.

5. Y. Akgul, D. Puschini, L. Vincent, M. Altieri, P. Benoit. Energy-efficient control through power mode placement with discrete DVFS and Body Bias. In *NEWCAS'15: 13rd IEEE International NEW Circuits And Systems conference*, Grenoble, France, 2015.

Patents

1. M. Altieri, S. Lesecq, E. Beigne, O. Heron. "Method and device for estimating circuit aging". FR Patent Application No. 1752811, Unpublished (filing date Mar. 31, 2017).

List of Figures

List of Tables

Titre : Estimation de la performance des circuits numériques sous variations PVT et vieillissement

Résumé

La réduction des dimensions des transistors a augmenté la sensibilité des circuits numériques aux variations PVT et, plus récemment, aux effets de vieillissement, notamment BTI et HCI. De larges marges de sécurité sont donc nécessaires pour assurer un fonctionnement correct du circuit, ce qui entraîne une perte d'énergie importante. Les solutions actuelles pour améliorer l'efficacité énergétique sont principalement basées sur des solutions de type ≪Adaptive Voltage and Frequency Scaling (AVFS)≫. Cependant, ce type de solution ne peut anticiper les variations avant qu'elles ne se produisent. Cette approche doit donc être amélioré pour traiter les problèmes de fiabilité liés au vieillissement. Cette thèse propose une nouvelle méthodologie pour générer des modèles simplifiés pour estimer la fréquence maximale du circuit f_{MAX}. Un premier modèle est créé pour estimer le délai de propagation du (des) chemin(s) critique(s) en fonction des variations PVT. Les effets BTI et HCI sont ensuite modélisés via une modification des paramètres du premier modèle. Construit à partir des modèles au niveau transistor, le modle de vieillissement obtenu prend en compte tous les facteurs qui influent sur le vieillissement, à savoir, la topologie des circuits, l'application, la tension et la température. La méthodologie proposée est validée sur deux architectures en technologie $28nm$ FD-SOI. Les modèles peuvent être alimentés par des moniteurs de température et de tension, ce qui permet une évaluation précise de l'évolution de f_{MAX}. Toutefois, ces moniteurs sont sensibles au vieillissement. Aussi, une méthode de recalibrage pour compenser les effets du vieillissement a été développée pour un moniteur numérique de température et de tension. Des exemples d'applications en ligne sont donnés. Les modèles sont également utilisés pour simuler des circuits complexes sous des variations de vieillissement, par exemple un circuit multi-cœur et un système AVFS. Cela permet d'évaluer différentes stratégies concernant la performance, l'énergie et la fiabilité.

Mots-Clés: fiabilité des circuits numériques, vieillissement, BTI, HCI, PVT, variabilité, architectures adaptatives

Title: Digital circuit performance estimation under PVT and aging effects

Abstract

The continuous scaling of transistor dimensions has increased the sensitivity of digital circuits to PVT variations and, more recently, to aging effects such as BTI and HCI. Large voltage guard bands, corresponding to worst-case operation, are thus necessary and leads to a considerable energy loss. Current solutions to increase energy efficiency are mainly based on Adaptive Voltage and Frequency Scaling (AVFS). However, as a reactive solution, it cannot anticipate the variation before it occurs. It has, thus, to be improved for handling long-term reliability issues. This thesis proposes a new methodology to generate simplified but nevertheless accurate models to estimate the circuit maximum operating frequency f_{MAX}. A first model is created for the modelling of the propagation delay of the critical path(s) as a function of PVT variations. Both BTI/HCI effects are then modelled as a shift in the parameters of the first model. Built on the top of device-level models, it takes into account all factors that impact global aging, namely, circuit topology, workload, voltage and temperature variations. The proposed modelling approach is evaluated on two architectures implemented in $28nm$ FD-SOI technology. The models can be fed by temperature and voltage monitors. This allows an accurate assessment of the circuit f_{MAX} evolution during its operation. However, these monitors are prone to aging. Therefore, an aging-aware recalibration method has been developed for a particular VT monitor. Examples of on-line applications are given. Finally, the models are used to simulate complex circuits under aging variations such a multi-core circuit and an AVFS system. This allows the evaluation of different strategies regarding performance, energy and reliability.

Key-Words: digital circuit reliability, aging, BTI, HCI, PVT, variability, adaptive architectures

Laboratoire : CEA-Leti, MINATEC - 17 rue des Martyrs,
38054 Grenoble Cedex 9, France