TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Technische Elektronik

Monitoring Concepts for Degradation Effects in Digital CMOS Circuits

Nasim Pour Aryan

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs

genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr.-Ing. U. Schlichtmann

Prüfer der Dissertation:

1. Univ.-Prof. Dr. rer. nat. D. Schmitt-Landsiedel
2. apl. Prof. Dr.-Ing. W. Stechele

Die Dissertation wurde am 03.03.2015 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 13.07.2015 angenommen.

Abstract

For safety critical electronic systems which are integrated in advanced technology nodes, the fulfillment of high reliability requirements is an important issue. This work focuses on modeling and monitoring the impact of degradation mechanisms on the reliability of digital CMOS circuits. Thereby, the impact of negative bias temperature instability (NBTI) as a dominant aging mechanism in 65nm and 40nm CMOS technologies is regarded. During the design phase the effect of NBTI is estimated by a novel aging analysis tool on circuit level. The developed tool evaluates the impact of an aging induced threshold voltage shift on the performance of the circuit and evaluates the permanent part as well as the recoverable component of NBTI.

To evaluate the impact of degradation mechanisms on the circuit over its lifetime and also to predict the possible upcoming failures, the circuit is equipped with a monitoring system. Since aging of devices results in a performance reduction of digital circuits, timing properties of the circuits are observed by in situ monitors. If in the extracted timing information a delay increase is detected, this is an indication for the degradation level and thus the reliability status of the circuit. To be able to interpret the extracted information, timing is converted to a digital code by either a centralized approach or a decentralized approach. In the centralized approach, several in situ monitors extract the remaining timing slack of several paths under test. The remaining slack of these paths and the entire circuit is then measured by a time to digital converter. In the decentralized approach, in situ monitors provide the remaining slack of the equipped path by digital outputs. The resulting digital code is transferred to higher layers of abstraction to determine the remaining slack of the entire circuit and thus to diagnose the system reliability.

By observing the reliability status of the circuit under test, necessary countermeasures can be carried out. For instance by dynamically adjusting the operating parameters, e.g. supply voltage and/or clock frequency, large design guard-bands chosen for a mostly unrealistic worst case condition are reduced. In this work, potential stabilization of circuit characteristics by adaptation of operating parameters is investigated in detail. Thereby, two safety critical applications are equipped with the monitoring system. The quantitative evaluations by simulation and experimental data show the applicability and the benefits of the monitoring concept for highly reliable systems.

Acknowledgment

This work was supported by the German Federal Ministry of Education and Research (BMBF) under the grant number 16M3091F within the funding project RELY-Reliability of SoCs for Applications like Transportation, Medical, and Industrial Automation. The work was concluded during my work as a research assistant at the Institute for Technical Electronics (LTE) at the Technische Universität München.
Foremost I would like to thank my professor and PhD thesis supervisor Prof. Dr. rer. nat. Doris Schmitt-Landsiedel for generously providing me the opportunity to carry out my research in such an interesting topic. During all stages of this work she provided continuous support and guidance. Her advice on my research as well as on my career has been invaluable. Her deep knowledge in physical aspects as well as her tremendous experience in integrated circuit design was of great assistance in this work.
Furthermore, I would like to thank the design team in Infineon Technologies, especially I would like to express my gratitude to Mr. Georg Georgakos for high level of expertise and administrative support. Valuable discussions with Mr. Georgakos and his support in concept, test chip development and hardware measurement were indispensable. I also want to thank Stefan Drapatz for his technical and administrative support during this work.
I would like to thank my past and present colleagues at the Institute for Technical Electronics. Thank you all for sharing your knowledge as well as for creating a supportive, communicative and enjoyable working environment. My special thanks to Markus Becherer, Stephan Henzler, Martin Wirnshofer, Cenk Yilmaz, Leonhard Heiß and Irina Eichwald. I would like to thank my other colleagues, Christoph Werner, Simon Stark, Stephan Breitkreutz-v. Gamm, Elisabeth Glocker, Michael Lueders, Josef Kiermaier, Mihail Jefremow, Francesco Santoro, Sebastian Kiesel, Grazvydas Ziemys and Andrew Giebfried. Moreover, I would like to thank Silke Boche, Rainer Emling, Christoph Kloesters, Hans Mulatz, Karl Demmel, Wolfgang Pielock, Norbert Leyh, Uwe Penning, Werner Kraus, Andrea Merkle as well as Bettina Cutrupia, Marta Giunta and Lydia Thalau for all their tool, hardware and administrative support. I would also like to thank the students who did their internships, bachelor and master theses under my supervision.
I would like to thank Prof. Dr.-Ing. Steffen Paul and my colleagues at the Institute of Electrodynamics and Microelectronics in University of Bremen, with whom I had close collaboration for the development and the measurement of a test chip

within the project RELY.
I would like to thank Prof. Dr.-Ing. Walter Stechele for collaboration in this project and his interest in this topic that made me even more motivated and eager to succeed.
Many thanks to my parents and my brother Naser, who always supported me in the path I chose in my life. My family has always been the greatest motivation in my life. Always believing in me, they provided me with the best circumstances to go for whatever fascinated me most. I can never thank them enough for being so great and wonderful.
My especial thanks goes to Ali Bigdelou, who has always been supporting and encouraging me during the hardest times for this work.

Contents

List of Figures

List of Tables

1. Introduction

1.1. Motivation

For safety critical applications such as aeronautic, automotive and medical, reliability is a crucial design goal. In today's advanced integrated circuits, with technology shrinking new effects in terms of reliability are emerging. Due to higher circuit sensitivity and more complexity, risk of failures in the circuits is increased. Moreover, the presence of dominant aging mechanisms induced by bias temperature instability (BTI), hot carrier injection (HCI), and time-dependent dielectric breakdown (TDDB) can get more severe in advanced technologies [1, 2, 3].
Therefore, in new technologies the decrease of circuit performance over time due to aging effects is not negligible. Depending on the structure and the operating conditions, transistor aging may result in more than 20% speed degradation [4]. In the extreme case, the circuit cannot continue to operate below the specified error rates at the same supply voltage and clock frequency for which it was originally designed. In the state of the art design, to prevent delay faults during the lifetime, worst case guard banding approach is used. Worst case guard banding approach considers extra margins for operating parameters such as supply voltage and clock frequency. Moreover, it recommends up-sizing of critical devices to fulfill reliability requirements under all operating conditions. However, aging of devices is strongly dependent on the workload and not all devices are aged to the same level. In addition, not all circuits may experience the worst case stress conditions during their operation.
For applications with high reliability requirements the reliability assessment is necessary during both the design phase and the lifetime of the circuits. Precise characterization of circuits during the design phase enables predicting the product lifetime [5]. Moreover, reliability simulations are necessary for developing new design methodologies for resilient and reliable circuits. To encounter the resulting increasing reliability costs and tighten the guard-bands, the reliability analysis in the early design phase is inevitable. For digital circuits flexible and yet sufficiently accurate analyzing methods and models on circuit level are required.
During the product lifetime the ability to predict near-future failures is highly advantageous for strict reliability requirements. Prediction of failures enables a reliability management during the entire lifetime of the circuit. Moreover, prediction of failures allows to reduce the margins and thereby reduces the cost and

ensures that the application specific reliability requirements are met. In dynamic approaches, the design is equipped with special circuitry, i.e. test structures that monitor the status of circuit-level properties. These test structures evaluate the impact of degradation mechanisms over the lifetime [6]. Therefore, in future highly reliable SoCs, internal system states will have to be constantly monitored. This provides a means for dynamic adaptation of the system to the current operating conditions. By dynamically adjusting the operating parameters, over-constrained guard-bands can be reduced resulting in area and power efficient design [7, 8, 9]. The gradual nature of aging mechanisms such as NBTI enables monitoring the degradation level [10] of digital circuits during lifetime to predict performance failures. Therefore, by monitoring the current reliability status of the circuit maintenance before failure or adaptation of operating parameters is possible.
This thesis addresses the circuit level reliability simulation considering NBTI aging mechanism during the design phase and proposes a reliability assessment approach for the circuits during the product lifetime.

1.2. State of the Art

Traditional modeling approaches for simulation of aging mechanisms in digital circuit blocks include transistor level simulation and gate level simulation [11, 12, 13]. Moreover, there is a lot of ongoing research on physical device models [14, 15, 16, 17]. The challenge is to utilize new advanced device models and bring them into the design flow by applicable methods for aging simulation of digital modules. The modeling approach presented in this work builds a bridge between advanced physical device models and reliability assessment of digital circuits in terms of aging.
There are several approaches that monitor the timing of the circuits, either as stand-alone circuits or as in situ monitors. In [18], the aging monitor circuit emulates the operating conditions of the critical circuits (replica). The monitor evaluates the aging of performance measures, and alerts the system if the degradation exceeds a certain pre-determined limit value. In another approach, tunable replica circuits (TRCs) [19], containing different logic stages, are implemented adjacent to each pipeline stage and calibrated to track local critical path delays. The authors in [20] use a buffer ring with a simple ring oscillator circuit and a pulse counter which monitors its oscillation period. The approach presented in [21] measures the beat frequency of two ring oscillators, one stressed and the other unstressed. However, dominant degradation mechanisms such as NBTI are mainly dependent on the workload and operating condition of the circuit. Since global monitors work with different operating condition and workload compared to the functional circuit, they are not able to provide completely accurate aging measurements. Moreover, within-die variations of process, voltage and temperature

affect the timing of such global speed monitor differently compared to the real circuit, which leads to inaccurate timing measurement. Therefore, in order to be able to predict the circuit failure accurately, using in situ monitors [9, 22, 23] is more advantageous than replicas and stand-alone circuits. In [24] a logic-level circuitry for detection of late-transitions occurring due to transistor aging in modern FPGAs is presented.
Nevertheless, there is a lack of highly accurate and aging resistant monitoring systems which evaluate the degradation level by a high resolution and report about the near future reliability status of the circuit.

1.3. Contribution of this Work

For the design process, this work proposes a novel reliability assessment tool which characterizes circuits precisely in terms of reliability during the design phase of digital circuits. The tool includes NBTI degradation and its recovery effect. The NBTI simulation tool is based on single device models and the corresponding measurement data. Thus, the tool builds a bridge from the device models to aging simulations of complex digital circuits. The circuits under test can be custom designed on transistor level and/or designed on gate level. Considering not only the permanent component of NBTI but also the recoverable part, the tool provides a useful means to prevent an early circuit failure at minimum costs.
For reliability assessment and management during the lifetime of the circuits, this work proposes accurate monitoring of the reliability status of digital circuits by measuring the remaining timing slack of the system. The optimized design and the implementation of the required aging resistant circuitry in 350nm, 65nm and 40nm is proposed and evaluated. Besides the quantitative evaluations regarding the accuracy and robustness of the monitoring circuitry, the overhead of the monitoring system is also evaluated. The applicability of the proposed approach is tested and the benefits and the efficiency of the monitoring system is demonstrated.

1.4. Structure of the Thesis

Chapter 2 gives an overview of the aging mechanisms in nanometer technologies and discusses the effect of NBTI as a dominant aging mechanism. In chapter 3 the developed NBTI simulation tool is explained in detail. Moreover, the reliability analysis utilizing the developed aging tool for test circuits is shown. Chapter 4 discusses the developed in situ timing monitoring systems. Chapter 5 shows the design of the required circuitry for the in situ monitoring system. In this chapter, the required circuitry for two timing slack measurement approaches, namely centralized and decentralized, are discussed. Chapters 6 and 7 give an evaluation of

the efficiency of the proposed approach in two different applications, respectively.

2. Aging and its Impact on Digital CMOS Circuits

Design technology faces challenges such as increasing silicon and system complexities [25]. This is due to effects such as the impact of process scaling and the increasing transistor counts, respectively. In advanced technology nodes the effect of time-dependent variations on device parameters over the lifetime has become crucial. In such technologies, the feature size is scaled more aggressively than the supply voltage. Thus, the electric field is enhanced, the impact of aging effects is increased and the parameter drifts are elevated. The shift in the device parameters results in overall performance degradation during the lifetime. To ensure an operation with an acceptable error rate below the defined value for the product lifetime, excessive margins to maintain the circuit reliability are necessary. However, the increasing reliability costs (including increased design cost, area, power etc.) tend to compensate the performance gain by moving from one technology node to the next. Consequently, a transition to the next technology node with the state of the art solutions might no longer be profitable. It should be noted that on the other hand an optimistic design might degrade the reliability of the circuit and result in wear out before the end of the expected product lifetime. This indicates a need for new low cost, reliable and resilient design methodologies. Developing new reliable design methodologies as well as defining guard bands for the worst case design requires new tools that estimate the reliability of the circuit in the early design phase.
In this chapter first the dominant aging mechanisms in nanometer technologies are discussed. Afterward, the effect of NBTI as a dominant aging effect in digital circuits is shown.

2.1. Aging Mechanisms

In this section the dominant wear-out mechanisms in advanced technologies such as bias temperature instability (BTI), hot carrier injection (HCI) and Time-dependent gate oxide breakdown (TDDB) are discussed.

2.1.1. Negative Bias Temperature Instability

One of the dominant degradation mechanisms in advanced CMOS technologies is bias temperature instability (BTI). BTI tends to increase the magnitude of the threshold voltage for metal-oxide-semiconductor field effect transistors (MOSFETs). Accordingly, the drain current is reduced and the delay of logic gates is increased. The major form of BTI mechanism is the negative bias temperature instability (NBTI) which occurs in PMOS transistors with negative gate source voltage. Due to non-constant field scaling, thinner gate oxides and the introduction of non-nitrided gate oxides the degradation caused by NBTI is increased [26]. NBTI is a charge trapping mechanism of defects in the gate stack when a transistor is under stress. When a PMOS transistor is under stress, a captured positively charged defect increases the magnitude of the transistor threshold voltage. This results in a decreased drive current and a temporal performance degradation of a gate. The threshold voltage drift is accelerated at higher negative gate-source voltages and elevated temperatures.
The stress condition for NBTI occurs while the transistor works in the triode region, i.e. at high $|V_{GS}|$ ($V_{GS} \approx -V_{DD}$) and low V_{DS} ($V_{DS} \approx 0$) voltages. In the triode region the channel is in inversion and can contribute to charge trapping. Hence, defects in the oxide can trap holes from the channel. Every defect possesses a specific capture time constant, τ_c depending on defect type and location in the oxide, determining when a certain defect becomes positively charged. Removing

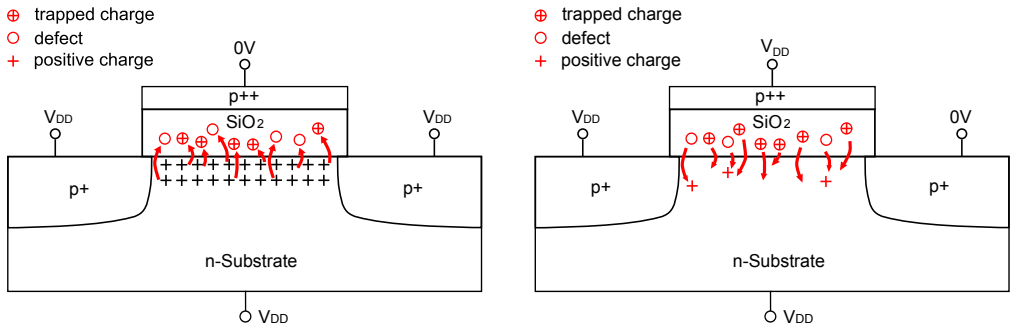


Figure 2.1.: Charge trapping and detrapping mechanism

the stress voltage leads to emission of a certain percentage of trapped charges with a typical emission time constant τ_e, which is different for every defect. Therefore, every defect can be individually characterized and identified by its stochastic capture and emission time constants. Fig. 2.1 shows a cross section of a PMOS during device stress phase. Defects with emission time constants smaller than the relaxation interval can return to their pre-stress occupancy. Activated defects with longer recovery time constants do not return to their uncharged state. Therefore,

the effect of these defects is visible as a permanent threshold voltage shift. Thus, for a given relaxation time only a certain percentage of charges will be emitted. This results in two distinct components contributing to NBTI, a removable and a permanent component [27].

2.1.2. Positive Bias Temperature Instability

In NMOS transistors the corresponding BTI phenomenon is positive bias temperature instability (PBTI). Similar to NBTI, PBTI originates from filling of preexisting electron traps in the oxide in combination with a process of trap generation [28]. However, PBTI appears with much lower magnitudes compared to NBTI if no high-k metals are used [29]. High-k oxides and metal gates are introduced to allow further gate oxide scaling with reduced gate leakage [30, 25]. In high-k oxides, the vertical field decreases but more traps are available to be charged. Therefore, with the introduction of high-k metal gates also PBTI becomes a significant issue and can no longer be ignored.

2.1.3. Hot Carrier Injection

The hot carrier injection (HCI) degrades the transistor due to accelerated carriers. HCI degrades both NMOS and PMOS transistors [10] and increases the effective channel resistance by degrading carrier mobility and increasing the magnitude of the threshold voltage. The term "hot" refers to the required carrier velocity corresponding to the average kinetic energy. For NMOS transistors damage occurs when in the lateral electric field the carriers gain enough energy to overcome the potential barrier between the silicon and the gate oxide and leave the channel. For PMOS transistors a similar mechanism occurs by hot holes. However, due to the lower mobility of holes compared to the electrons, HCI is more critical for NMOS devices. Fig. 2.2 shows the HCI mechanism in NMOS devices as explained by the lucky electron model [31, 32]. When the NMOS transistor is turned on, impact ionization results in generation of hot carriers in the high electric field close to the drain. Some of the hot carriers are injected into the gate. Hot carriers degrade the dielectric and shift the threshold voltage of the transistors. In the contrary, holes are repelled from the drain and gate, resulting in a substrate current.
Similar to NMOS transistors, in PMOS devices a threshold voltage increase is induced by the small portion of carriers which are caught in the gate oxide. A reduced drain current by HCI decreases the gate performance.
In scaled technologies other mechanisms such as electron-electron scattering (EES) [33] or multiple vibrational excitation (MVE) [34] give a different explanation of HCI process. The rate of hot carrier degradation is related to the length of chan-

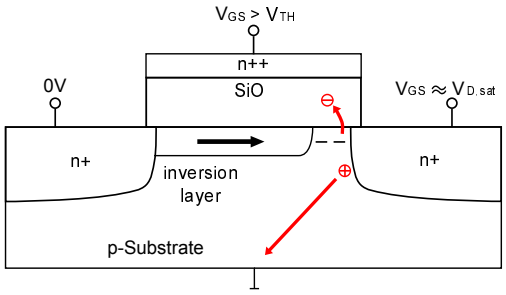


Figure 2.2.: Hot carrier injection mechanism

nel, oxide thickness and the supply voltage. An enhanced HCI degradation can be observed due to a continuous increase in the lateral electric field in scaled technologies. Regarding digital CMOS circuits, in contrast to NBTI, HCI degradation requires a flow of current and thus occurs during transitions in between logic states.
HCI degradation strongly depends on the signal slope described by parameters such as the slew rate or the fan-out [35]. HCI is negligible in the fast slope regime and becomes nearly as severe as NBTI for slow input slew rates. In high-k devices PBTI is believed to have a larger influence than HCI [36].

2.1.4. Non-Conducting-HCI

Non conducting HCI (NCHCI) is dominant in NMOS transistors in the non-conducting state. NCHCI occurs due to the leakage currents. NCHCI is becoming more of a problem in the nanometer technologies as leakage currents increase by scaling. In this mode, due to a high drain-source voltage, hot carrier injection occurs in the gate-drain overlap region and generates interface traps which degrade the current. However, parameter shifts due to conducting HCI degradation are dominant compared to NCHCI and lead to a higher performance degradation.

2.1.5. Time-dependent Dielectric Breakdown

Time-dependent dielectric breakdown (TDDB) of SiO_2 occurs when the gate oxide breaks down as a result of a conducting path through the oxide to the substrate. When a large electric field is applied to the dielectric which is larger than the maximum it can sustain, a hard breakdown occurs. When lower electric fields are applied, the wear-out mechanism of the insulator occurs more slowly and finally results in the time-dependent dielectric breakdown [37].
Due to the scaling of the interconnect dimensions and using low-k materials,

TDDB is becoming more of a problem [38].

2.2. Technology Scaling

To increase the performance and the density by technology scaling the electric field needs to be kept constant. If the voltage levels are scaled proportional to the device dimensions and the doping is scaled inversely proportional to the device dimensions to obtain the required threshold voltage, the resulting electric field would stay constant. The constant electric field then leaves the carrier velocities unchanged. Ideally, if the dimensions are scaled by $1/\lambda$ ($\lambda > 1$), the circuit is sped up by a factor of λ. Moreover, the power consumption is reduced by a factor of $1/\lambda^2$ due to the voltage scaling.
However, it is not possible to scale the supply voltage by $1/\lambda$ due to the system requirements, small noise margins and etc. Scaling down the supply voltage to higher values results in an increased electric field. This results in effects such as hot carrier injection which degrade the MOSFET transistors operation [39].
To prevent loss of performance when scaling down the supply voltage, the transistor threshold voltage needs to be decreased. This in turn results in higher off-state leakage currents and increases the power consumption. To reduce the leakage currents and achieve lower power consumption high-k materials have been introduced, which increase the sensitivity to PBTI.
Scaling the oxide thickness by few atomic layers increases the effect of NBTI and TDDB. Moreover, introduction of nitrides in oxides increases the sensitivity to NBTI [40].

2.3. Impact on Digital Circuits

This work considers the effect of NBTI as a dominant aging mechanism in advanced digital CMOS circuits. NBTI is best modeled as a voltage drift corresponding to the ΔV_{th} at the gate of a PMOS transistor. In this section the effect of an aging induced threshold voltage drift due to NBTI on the performance of the standard library components is discussed. As long as PBTI is neglected, only the degradation of the threshold voltage of the pull-up stage of a CMOS gate affects the performance of the digital circuit. To determine the performance degradation due to NBTI, the correlation of the gate delay to an increased $|V_{th}|$ is discussed. Moreover, the conditions for the occurrence of the threshold voltage shift are evaluated.

2.3.1. Logic Gates

In this part the effect of NBTI on combinatorial gates in a commonly used static CMOS logic is discussed. In CMOS logic every combinatorial gate consists of a pull-up and a pull-down network. The gate propagation delay is defined as the time from a $V_{DD}/2$ point of an input transition to a $V_{DD}/2$ point of the output transition. Since the pull-up network is only composed of PMOS transistors and the pull-down stage only of NMOS transistors, it is sufficient to consider only the degradation of the pull-up stage for NBTI. Here, each logic stage is considered independent of its pre and post stages.
NBTI is strongly dependent on the applied stress pattern. Transistors in the same circuit might not experience the same stress due to a varying stress pattern. Whether the PMOS transistor is under stress or not depends on the applied input patterns and the structure of the circuit, e.g. if a PMOS stacking is present. Moreover, for multi-stage gates, which are composed of several single-stage gates in series additionally all internal nets connected to the transistor gate terminals are considered. Thus, different circuit structures and workloads result in different shifts in the threshold voltage of PMOS devices. Different threshold voltage shifts result in different increases in delays for different gates. This asymmetric delay degradation leads to parts of a circuit being more prone to aging than others. The following describes dependency of NBTI on the circuit structure.
For an inverter the stress condition is fulfilled when the PMOS is negatively biased with respect to the source and drain. This means that a logic "0" at the gate terminal is applied, resulting gate source voltage is $V_{GS} = -V_{DD}$ and the PMOS is degraded. Since the transistor is in inversion, the drain voltage is also charged to V_{DD} ($V_{DS} = 0$). For the sake of simplicity the small voltage drop over the transistor itself caused by its on-resistance is neglected.
For a falling input transition the degraded PMOS results in a degraded output slope as the output load is charged slower. Consequently, the gate delay is increased. For the rising input transition of an inverter the propagation delay change can even be negative. The reason is that the NMOS pull-down discharges the output load faster due to the weakened PMOS. However, this effect is only observed for a slow input slew rate.
The delay for an inverter can be approximated by the α-power law for the short-channel MOS transistors and derived as [41]

$$t_{pHL} = \left(\frac{1}{2} - \frac{1-\nu_T}{1+\alpha}\right) \cdot t_T + \frac{C_L\, V_{DD}}{2\, I_{D0}}, \qquad \nu_T = \frac{V_{thp}}{V_{DD}} \tag{2.1}$$

where I_{D0} is the drain current at $V_{GS} = V_{DD}$, V_{thp} is the threshold voltage of the PMOS and t_T is the transition time of the input waveform. For a typical short-channel MOS transistor α is approximately 1. It can be observed that a higher threshold voltage results in an increased gate delay. Moreover, the delay is

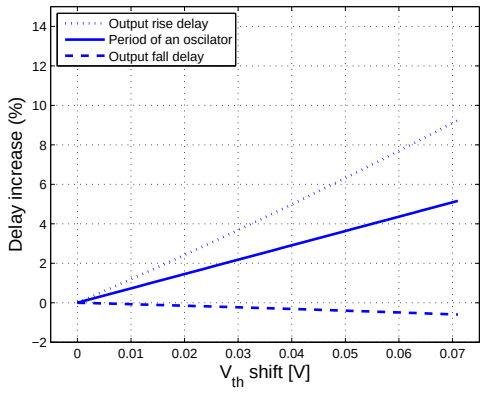


Figure 2.3.: Simulated dependence of the inverter delay on the threshold voltage shift

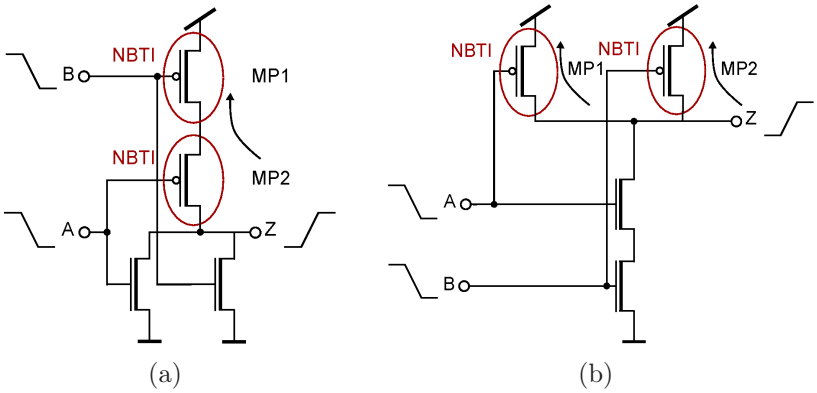


Figure 2.4.: NBTI degradation of the 2-input a) NOR gate and b) NAND gate

dependent on the input slope by t_T. Figure 2.3 shows the simulated dependence of the inverter delay on the threshold voltage shift.

For multiple input gates the applied input pattern determines if a PMOS transistor experiences NBTI stress or not. This is the case if a multi-input gate contains transistors connected in series, i.e. in a stack. As an example, Fig. 2.4a shows a 2-input NOR gate where two PMOS transistors are in a stack. For transistor MP1, similar to an inverter, if a logic "0" is applied at input A, the PMOS is under stress. However, for transistor MP2 to enter the stress condition, both inputs need to be "0". As a consequence, either both PMOS transistors are experiencing NBTI stress or just the one connected directly to V_{DD}. Hence, whether the second transistor in series is under stress depends on the input combination [42, 13]. For gates with more transistors in a stack even more cases need to be considered.

However, not for all multi-input gates the states of other transistors need to be

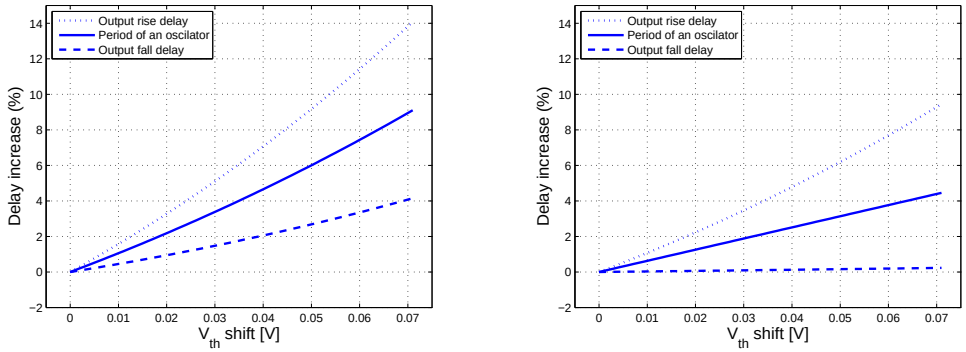


Figure 2.5.: Simulated dependence of the 2-input a) NOR gate and b) NAND gate delay on the threshold voltage shift

regarded. In comparison to the NOR gate, in a 2-input NAND gate (see Fig. 2.4b) the sources of all PMOS transistors are tied to the supply voltage V_{DD} and hence V_{GS} is always either $-V_{DD}$ or 0. Therefore, if the input signals are independent, the stress condition for each PMOS is also independent of the others.
It can be assumed that the delay of a gate in dependence of V_{th} based on the α-power law follows [43]

$$\tau_d \approx \frac{C_L\, V_{DD}}{I_{D0}} \propto \frac{V_{DD}}{(V_{DD} - V_{th})^{\alpha}} \tag{2.2}$$

Where the the exponential parameter α is between 1 and 2. From this equation, an expression for the gate delay degradation with the threshold voltage can be formed [44]

$$\frac{\Delta\tau_d}{\tau_d} \propto \frac{\alpha\, \Delta V_{th}}{(V_{DD} - V_{th})^{\alpha}} \tag{2.3}$$

Figures 2.5a and 2.5b show the simulated dependence of the gate delays on the threshold voltage shift for NOR and NAND gates, respectively. For the NOR gate, an increased performance degradation compared to the NAND gate is observed. This is due to stacking of PMOS transistors in the NOR gate. Thus, degradation of two PMOS transistors in series results in a higher delay increase compared to the NAND gate.

2.3.2. Flip-Flops

Synchronous digital designs use edge-triggered flip-flops as the sequential cells. Therefore, the impact of a threshold voltage drift on commonly used flip-flops is

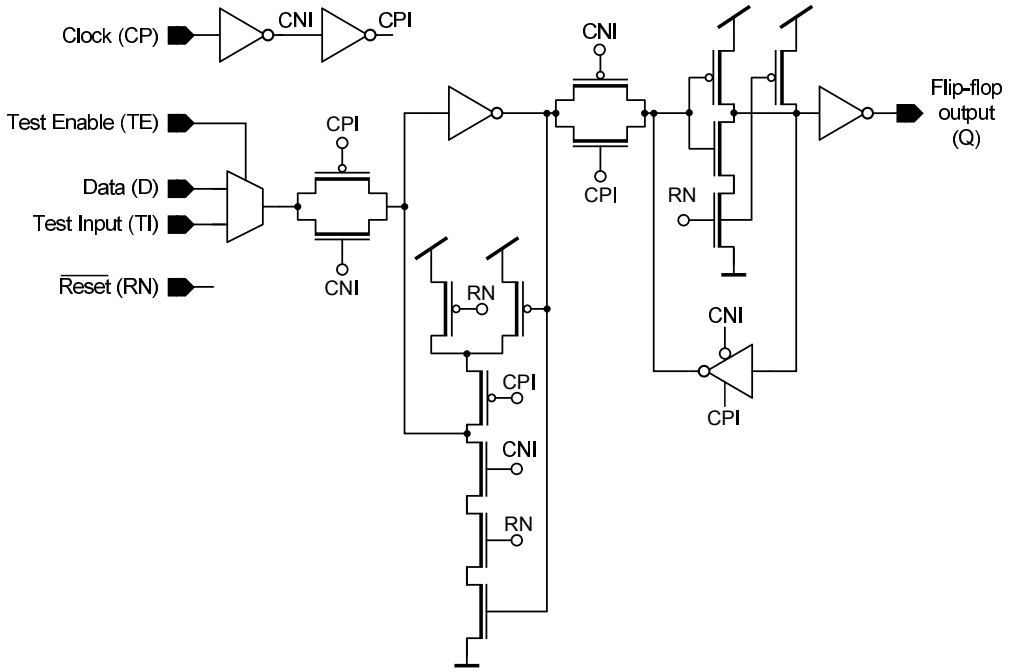


Figure 2.6.: The state of the art scan master-slave D-flip-flop with asynchronous reset and multiplexed data and test inputs

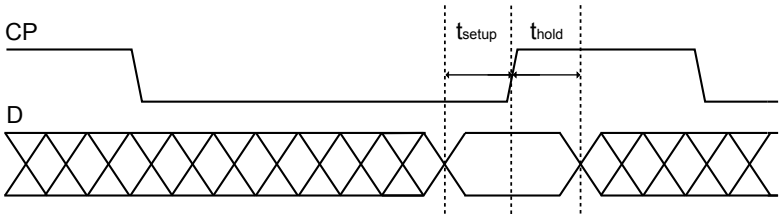


Figure 2.7.: Timing figures of flip-flops including the setup time and hold time

investigated. Here, the setup and hold times of flip-flops should not be violated. To determine the impact of NBTI degradation on flip-flops an state of the art scan master-slave D-flip-flop is evaluated. Therefore, the setup time degradation of a the flip-flop with asynchronous reset and a multiplexer as input stage (as in Fig. 2.6) is evaluated. The setup time constraint will be violated as soon as the path delay is increased such that the data signal arrives after the setup time at the flip-flop (see Fig. 2.7). Thus, an increased setup time by aging might result in setup time violation and timing errors. However, in high performance systems, the setup time amplification is not significant compared to the large delay of the logic gates.
Fig. 2.8 shows a simulation of a fresh and an aged flip-flop for three corner cases and a supply voltage range from $V_{DD} = 0.9\text{V}$ to $V_{DD} = 1.2\text{V}$. The flip-flop was

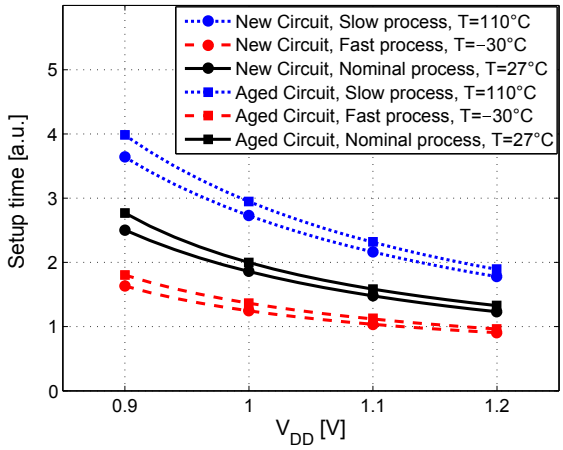


Figure 2.8.: Setup time comparison of a fresh and an aged master slave D-flip-flop

aged in AC mode for 10 years at a temperature of 85 °C with the developed aging tool which is introduced in the next chapter. As expected, the setup time of the aged flip-flop is increased compared to the fresh one. Fig. 2.8 shows an increase between 7.8% to 10.5% depending on the supply voltage. For the nominal process at a supply voltage of $V_{DD} = 1.2$V the setup time displays an increase of only 7.8%.
For the hold time constraint the short combinatorial paths are considered as the hold time is violated if a signal is not stable long enough to be captured correctly. An increased hold time compensates the increased gate delay along a path to a certain amount. Hence, a hold time degradation has no crucial impact and is thus negligible.

2.4. Aging and Digital Design Flow

To meet the application specific reliability requirements, the traditional approach in the digital design flow introduces sufficient safety margins by the worst-case guard banding approach. Hence, very conservative safety margins are encountered in scaled technologies where drift-related parameters severely impact the circuit performance. However, generous guard-bands result in a waste of power, performance and area. Since aging effects such as NBTI lead to an increase the minimum operating voltage and reduce the maximum operating frequency, variation-aware design techniques are more advantageous to ensure reliable products.
To achieve a high level of system reliability, all aging effects need to be taken into account. It should be noted that the impact of device degradation on the circuit performance cannot be generally predicted but is strongly dependent on the op-

erating conditions [45]. This work only focuses on the modeling and simulation of NBTI as the most dominant effect in advanced technologies.

2.5. Aging and IC Qualification

To evaluate the product lifetime accelerated conditions are used to imitate the expected reliability and operating lifetime over shortened test period [46].
Figure 2.9 shows the failure rate as a function of the product lifetime. As can be seen in this picture, the number of detected defects in the chip is relatively large during the infant mortality period. To evaluate this phase in terms of reliability, the burn in tests are designed which accelerate the aging by 1.3 to 1.4 times the operating temperature and voltage [47]. Depending on the application, different burn-in tests such as DC, dynamic (by applying input vectors), test in burn-in etc. are performed. Another reliability test is the high temperature operation life test

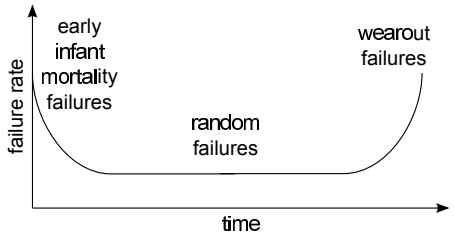


Figure 2.9.: The bathtub curve displaying the failure rate as a function of the product lifetime

(HTOL) which applies accelerated temperature and voltage over longer periods, e.g. 1000h. HTOL is performed over small sample sizes to determine the lifetime and the failure rate of the products [46]. According to the typical foundry IC qualification methods ([48]), HTOL is performed by an accelerated voltage of 1.1 to 1.2 times the nominal value, and a stress temperature of $Temp = 125\,^{\circ}\mathrm{C}$ or higher. The passing criterion of the HTOL test is determined by the required failure rate. Based on the HTOL acceleration model and the Arrhenius relationship, various points in the lifetime of the devices with different failure rates can be mapped to different stress intervals [49].
In the accelerated setup utilized in this work to identify the degradation status of the circuit over the lifetime, the aging of chips is accelerated with alternating stress measurement phases. Therefore, it would be possible to track the gradual aging of devices.

2.6. Summary

In this chapter the dominant aging mechanism in advanced technologies were discussed. Moreover, the effect of scaling on the aging effects was explained. Afterward, the impact of a threshold voltage shift induced by NBTI as a dominant aging on digital circuits was discussed. Finally ensuring the product reliability during the design phase and IC qualification for accurate estimation of the lifetime reliability was discussed.

3. Aging Simulation in Digital CMOS Circuits

In safety critical applications precise characterization of circuits to predict the lifetime reliability is a key challenge. This chapter proposes a reliability assessment tool to model and simulate the NBTI degradation including its recovery effect during the design phase of digital circuits.
Since the 90nm node, NBTI is considered as the predominant cause of device reliability degradation and lifetime limitation. To deal with reliability concerns, state of the art design introduces generous safety margins arising from the worst-case scenario. However, systems used in space, avionic, and biomedical applications require high reliability levels. Therefore, in such applications very conservative safety margins are used, where drift-related parameters severely impact the circuit performance. This results in high waste of power, area and performance [50]. Nevertheless, device degradation is strongly dependent on the circuit structure and workload. Therefore, all devices within a circuit are not aged to the same level. Moreover, for reliability issues with a recovery effect such as NBTI, after the stress is removed, the drift in parameters is partly recovered [51, 16, 52]. As many guard banding approaches consider only the permanent component of degraded parameters under worst case stress, they underestimate the recoverable component of NBTI.
Therefore, accurate prediction of the NBTI effect is still challenging due to its strong recovery characteristic. To achieve this, not only the workload and structure of the circuit, but also the stress and recovery behavior of aging effects need to be taken into account.
Despite of the known existence of the reliability challenges, there is still a lack of flexible and yet sufficiently accurate analyzing methods on circuit level for digital circuits. Therefore, this chapter proposes to predict the timing degradation due to NBTI induced aging with a novel sufficiently accurate model that is capable of analyzing even complex digital circuits when considering both NBTI stress and recovery.

3.1. Conventional NBTI Modeling

3.1.1. Physical Single Device Model

The charge trapping models for NBTI analysis assume that under stationary conditions defects randomly exchange charge with the substrate. Considering the charge trapping/detrapping mechanism, every defect can exist in two states, a charged (positive) and an uncharged (neutral) state with stochastic transitions between them determined by two time constants. Here, a two state Markov model can explain the charge trapping detrapping mechanism, as shown in Fig. 3.1. Therefore, every defect has a specific capture time constant τ_c determining when

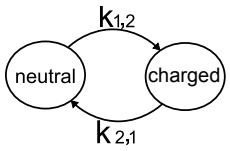


Figure 3.1.: Two-state Markov model for defects in the oxide of the transistor

a certain defect becomes positively charged. Removing the stress voltage leads to emission of a certain percentage of trapped charges with a typical emission time constant τ_e. Mean values of the transition probabilities between neutral and charged states in the Markov model can be expressed as

$$\overline{\tau_{charged \rightarrow neutral}} = \overline{\tau_c} = E\{\tau_c\} = \frac{1}{k_{1,2}} \tag{3.1}$$

$$\overline{\tau_{neutral \rightarrow charged}} = \overline{\tau_e} = E\{\tau_e\} = \frac{1}{k_{2,1}} \tag{3.2}$$

in which the transition rate $k_{i,j}$ is the transition probability per unit time.
If the stress is applied, meaning that the V_{SG} bias is switched to a larger value, the bias dependent capture time constants become smaller. This results in defects with smaller capture than emission times to be in their charged states. Capture and emission time constants are widely distributed over time (due to different properties and location of the defects) and transitions occur at different times for each defect. Both capture and emission time constants range from the microseconds regime up to several years. As mentioned before, a captured positively charged defect increases the magnitude of the transistor threshold voltage. The summation over all defects in turn results in the overall NBTI degradation.
Recent investigations have shown that the charge exchange between the border states and the channel occurs via a nonradiative multiphonon process (NMP). NMP has also been exploited for random telegraph noise [53] and $1/f$ noise in devices [54]. The large range of time constants described by the NMP results

from the thermal barrier upon charge capture.
From a two-state Markov model and the nonradiative multiphonon theory, the two time constants describing the capture and emission times of an individual defect can be derived [27]. After the derivation of the stochastic process which describes the time constants of the charge trapping and detrapping, the transmission rates are linked to a physical model. The nonradiative multiphonon theory renders the best fit with the experimental data. The basic idea of NMP comes from the premise of the conservation of the total energy, meaning that charge capture and emission are only possible if the sum of the electronic and the vibrational energies of a defect are conserved. Figures 3.2a and 3.2b show the adiabatic defect potentials for a two state defect versus its reaction coordinate. Only nonradiative transitions are considered. Thus, no direct transition between

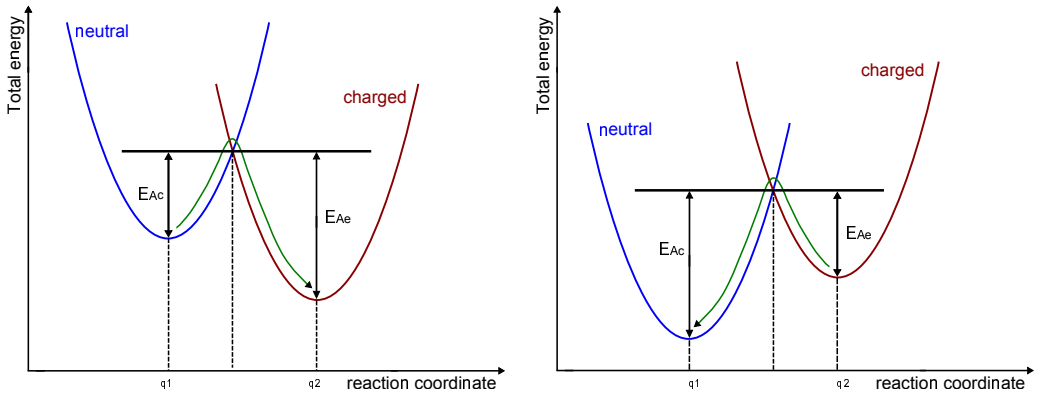


Figure 3.2.: Bias dependence of nonradiative multiphonon (NMP) transition rates due to a shift of the defect energy level for charge a) capture and b) emission mechanism

the states is possible. The energy increase is due to phonons and a transition is possible at the intersection point of the parabolas. Using Boltzmann statistics for the occupation probability, the transition rates can be calculated by a first order kinetic rate equation [55] and derived as

$$k_{12} \approx p\, v_{th}\, \sigma\, e^{-\beta E_{Ac}} \tag{3.3}$$

$$k_{21} \approx N_v\, v_{th}\, \sigma\, e^{-\beta E_{Ae}} \tag{3.4}$$

where $\beta = 1/(k_B\, Temp)$, in which $Temp$ is the absolute temperature and k_B is the Boltzmann's constant. σ is the capture/emission cross section and the thermal velocity is $v_{th} = \sqrt{8\, k_B\, T/(\pi\, m)}$. The density of holes is $p = N_V \exp(\beta(E_V - E_F)$ and N_V is the effective density of states in the valence band. E_V is the energy of the valence band and E_F is the Fermi energy. The barriers E_{Ac} and E_{Ae} are given

by the adiabatic potentials and determine the energies that have to be overcome for a change in charge state [27].
With the derived expressions for τ_c and τ_e from equations 3.1 and 3.2, respectively, as the time constants are uncorrelated to the depth of the defect into the oxide equations **3.3** and 3.4 can be simplified to

$$\tau_c = \tau_0 \, e^{\beta \, E_{Ac}} \tag{3.5}$$

$$\tau_e = \tau_0 \, e^{\beta \, E_{Ae}} \tag{3.6}$$

where τ_0 is an experimentally determined prefactor depending on the technology. To describe more than one individual defect the capture emission time (CET) map modeling of [56] is utilized. A CET map describes the distribution of capture and emission times of traps in the gate oxide. Experimentally gathered CET maps for most nitrided gates show that NBTI in general consists of two distinctly different components: A recoverable component R and a permanent component P. This results to the fact that not only the permanent component is the crucial degradation mechanism determining the lifetime [57], but also the recoverable part is highly crucial in transient reliability assessment. The reason is that in low activity nodes within the circuit toggling with much lower frequencies than the system clock, a temporal (recoverable) increase of V_{th} shift can be present. As this ΔV_{th} is not recovered to a great extent within the interval of few clock periods, the recoverable component of NBTI has to be considered.
The time constants for a specific trap are weakly correlated and depend on the gate bias voltage and the temperature. However, instead of using the time constants directly, the distribution of the activation energies can be modeled so that the parameters impacting them need not be regarded. The overall degradation S arises from adding the recoverable and the permanent components [58]. The recoverable and permanent components are well described by two regular bivariate normal distributions derived as

$$\Phi_R(E_{AR}, \mu_{R\,c,e}, \boldsymbol{C_R}) = \frac{1}{2\pi\sqrt{det(\boldsymbol{C_R})}} \exp(-\frac{1}{2}(E_{AR} - \mu_{R\,c,e})^T \boldsymbol{C_R^{-1}} (E_{AR} - \mu_{R\,c,e})) \tag{3.7}$$

$$\Phi_P(E_{AP}, \mu_{P\,c,e}, \boldsymbol{C_P}) = \frac{1}{2\pi\sqrt{det(\boldsymbol{C_P})}} \exp(-\frac{1}{2}(E_{AP} - \mu_{P\,c,e})^T \boldsymbol{C_P^{-1}} (E_{AP} - \mu_{P\,c,e})) \tag{3.8}$$

Here, P and R denote the permanent and recoverable terms for NBTI. $\mu_{c,e} = [\mu_c,\ \mu_e]^T$ is the vector of mean activation energies for capture and emission mechanisms. $\boldsymbol{C}$ is the correlation matrix for the capture and emission mechanisms in

the energy domain

$$\boldsymbol{C} = \begin{bmatrix} \sigma_e^2 & \rho_{c,e}\sigma_e\sigma_c \\ \sigma_e\sigma_c & \sigma_c^2 \end{bmatrix} \tag{3.9}$$

where σ_c and σ_e are the standard deviations of the activation energies for charge capture and emission mechanisms, respectively and $\rho_{c,e}$ is the correlation factor between the activation energies for charge capture and emission.
All parameters of the distributions in equations 3.7 and 3.8 can be obtained by a least-squares fit to the measurement data. Eventually, the threshold voltage shift results from the integration over all defect energies. The parameter shift increases with time under negative gate bias and at elevated temperature. The threshold voltage shift of the recoverable and permanent component are given by

$$\Delta V_{th,R}(t) = \left(\frac{V_s}{V_{s0}}\right)^m \cdot \left(\int_\alpha^\beta \int_\gamma^\delta \Phi_R(E_{AR}, \mu_{R\,c,e}, \boldsymbol{C_R}) \mathrm{d}E_{ARe} \mathrm{d}E_{ARc}\right) \tag{3.10}$$

$$\Delta V_{th,P}(t) = \left(\frac{V_s}{V_{s0}}\right)^m \cdot \left(\int_\alpha^\beta \int_\gamma^\delta \Phi_P(E_{AP}, \mu_{P\,c,e}, \boldsymbol{C_P}) \mathrm{d}E_{APe} \mathrm{d}E_{APc}\right) \tag{3.11}$$

where V_s is the bias voltage, and V_{s0} and m are experimentally determined constants [56]. α, β, γ and δ are the boundaries of integration, which will be discussed in the following. The term $(V_s/V_{s0})^m$ shows the strong dependence of ΔV_{th} on the bias voltage. The overall ΔV_{th} is then given by the sum of the two components

$$\Delta V_{th} = \Delta V_{th,R} + \Delta V_{th,P} \tag{3.12}$$

Based on the type of the stress (DC or AC-pattern), the stress/recovery times and the stress duty factor (DUF) determine the boundaries of the integration in equations 3.10 and 3.11. In case of a DC-stress, a constant stress voltage is applied to the transistor for a certain time t_{stress} after which the transistor enters recovery for the relaxation time t_{relax}. On the other hand, for AC-stress a periodic pulsed gate source voltage with a constant frequency f_{AC} and DUF is applied to the transistor. The area of integration in equations 3.11 and 3.10 to determine the threshold voltage shift for an AC-stress scenario is then dependent on the duty factor as well as the stress period $T_{AC} = 1/f_{AC}$. Compared to a DC-stress, an AC-stress leads to a more optimistic prediction of the degradation over the lifetime and thus less conservative safety margins.
Depending on the measurement technique used for generating capture emission time maps, only a certain part of the existing charged defects are seen in the

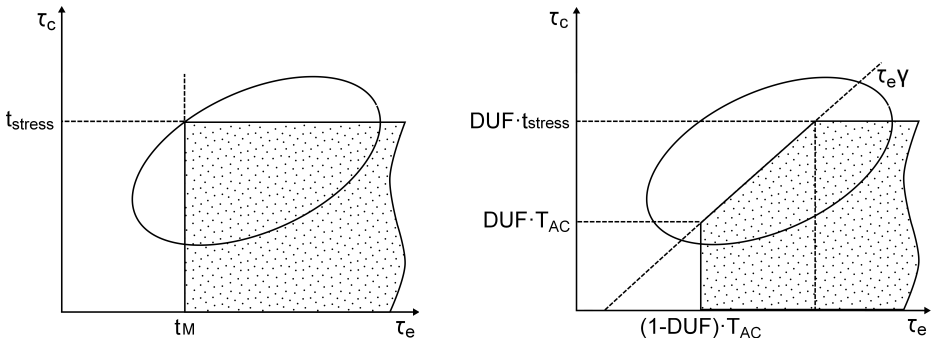


Figure 3.3.: Charged defects in the τ-domain for a) DC-stress scenario measured with "measure-stress-measure" method and b) AC-stress scenario ([56])

resulting ΔV_{th} [57]. Thus, only the contributing defects to achieve a fit of the calculated values and the experimentally obtained data need to be regarded. To specify the integration boundaries and to fit the model correctly the measurement technique is considered. Figures 3.3a and 3.3b show the charged defects energies in the τ-domain. The marked areas correspond to defects that are charged for the chosen measurement technique and thus contribute to the threshold voltage shift. The total area of charged defects in the τ_e-τ_c-plane results in the total charge. An integration over all defects then determines the ΔV_{th} [56].

As shown in Fig.3.3a, in the DC-stress scenario all defects with emission time constants smaller than the measurement delay t_M are discharged. All defects from 0 to t_{stress} and from t_M to ∞ are charged and contribute to the degradation. The ΔV_{th} for any combination of t_{stress} and t_{relax} in between this area is obtained from an integration over the bivariate normal distribution of the activation energies (equations 3.10 and 3.11).

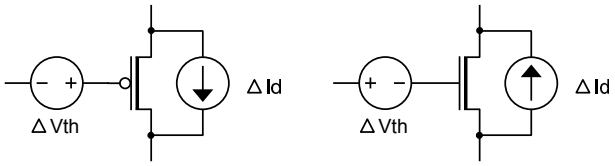
In the AC-stress scenario strong duty-factor dependence is observed and the area of integration is divided into three regions. Defects with emission times smaller than $(1 - DUF) \cdot T_{AC}$ are emptied during each relaxation cycle and recharged during the next stress cycle. Defects with very large emission times stay charged in the emission cycle and charge up to $\tau_c < DUF \cdot t_{stress}$. In the transition region in between, all defects with $\tau_e < \tau_c/\gamma$ with $\gamma = DUF/(1 - DUF)$ remain charged (see Fig. 3.3b.)

This can also be interpreted as integration over the whole region similar to Fig. 3.3a with a subsequent subtraction of the upper triangle, whose lower boundary is described by the linear function $\tau_c = \tau_e \cdot \gamma$. The missing upper triangle of the defect density map suggests a sequential filling of the defects. During the AC-stress, defects fill up from the bottom to the top while the occupancy level is

continuously rising with stress time. Hence, only the levels right of the linear function $\tau_c = \tau_e \cdot \gamma$ are filled. Interruption of stress leads to discharging of the defects beginning from the left to the right until the point $\tau_e = (1 - DUF) \cdot t_{stress}$ is reached. Due to the missing triangle the ΔV_{th} for AC-stress is much smaller. Therefore, even for very high duty factors a decreased threshold voltage shift is observed. Eventually, after a certain recovery time, the AC-stress contains a rectangle and the recovery curves obtained by an AC-experiment merge with the recovery curves of a DC-experiment.

3.1.2. Transistor Level Simulation

There are several approaches which analyze the impact of aging effects on transistor level [11, 12]. The degradation due to the aging effects such as NBTI, PBTI, CHCI and NCHCI on transistor level is modeled in terms of shifts in the transistor threshold voltage, ΔV_{th}, and drain current, ΔI_d [59, 60, 61]. NBTI and PBTI reliability mechanisms mainly increase the magnitude of V_{th} of MOS transistors and conducting and non-conducting HCI reduce the I_d. HCI also gives a smaller contribution compared to BTI to the V_{th} shift. Figure 3.4 shows modeled V_{th} and I_d shifts on transistor level.
Commercially available tools on transistor level such as BERT [62] or RelXpert [63] simulate the fresh circuit and store the current and voltage waveforms at the transistor terminals to determine the workload for each transistor. Afterward, degraded transistor models are generated and a second simulation with the aged circuit is performed. However, such models are only accurate if they describe the degradation of the transistors precisely. Moreover, they are proprietary and not interchangeable by the user. More importantly, such tools are mainly built for the analysis of small analog circuits and are not capable of verifying the timing constraints of more complex digital circuits with thousands of transistors and very different possible stress scenarios. Thus, a more resource efficient method is required for the aging simulation of digital modules.



(a) (b)

Figure 3.4.: Modeling the parameter shifts on transistor level for a) PMOS and b) NMOS devices

3.1.3. Gate Level Simulation

Gate level approaches evaluate the effect of aging on the digital circuit on the gate level. Gate level approaches based on aged look-up tables (LUT) store the gate performance as a function of the input signals, output load and operating conditions such as temperature, supply voltage and signal probability. These approaches characterize aged netlists for predefined operating conditions over lifetime. However, a very time consuming re-characterization is necessary for other operating conditions [45].
Other modeling approaches on gate level commonly estimate the aged gate delay [13] due to a threshold voltage shift as the sum of the delay of the fresh circuit and the aged delay. However, these models are inaccurate compared to device level models as only one threshold voltage shift per gate is considered and no equation for the aged output slope is provided. Gate level approaches such as [13] use a canonical gate model which provides the aged gate performance for parameter drifts of individual transistors. However, with gate level approaches analysis of custom designed circuits is not possible. Moreover, to efficiently update the aging analysis for new technologies it is advantageous to have interchangeable parameter models. This necessitates separating the physics behind the aging model from the aging analysis tool-set. Separating physical model from the aging framework enables to integrate effects such as recovery of aging mechanisms into the aging analysis tool. Moreover, it speeds up integrating new developed models for scaled technologies into the aging tool-set.

3.2. From NBTI Device Models towards Aging Assessment of Digital Circuits

In this chapter the NBTI simulation tool applicable on circuit level is discussed [5]. The developed approach combines gate level and transistor level approaches to accurately characterize digital circuits with low computational effort compared to the transistor level approaches. The approach performs an evaluation of the timing behavior of the digital circuit over the lifetime. The delay and output slope calculations result from a SPICE simulation of an aged netlist. As the physical model is separated from the circuit analysis, integration of updated models for new technologies is highly efficient. The developed aging analysis tool utilizes the NBTI device model and measurement data of [56], in which both NBTI stress and relaxation are regarded. In the circuit level modeling approach presented in this section the effect of the recoverable component of NBTI is carefully evaluated. Therefore, the tool is more accurate than current approaches for aging analysis of digital circuits.

3.2.1. Workload Definition

The model of [56] distinguishes DC- and AC-stress scenarios. Compared to a pure DC-stress, a shorter stress time is applied in the AC case and a number of defects can recover. This results in a lower threshold voltage shift depending on the duty factor. Hence, considering AC-stress instead of the worst-case DC-stress will result in a substantially lower prediction of the delay degradation and thus a reduction of the necessary guard-bands.
However, this assumed periodic stress with a certain frequency and DUF cannot represent signals in digital circuits. Instead, a model is required that deals with diverse aperiodic patterns. This is due to different activity rates of the nodes within the circuit. Since it is not feasible to track the signals for every PMOS to gather information about this aperiodic behavior, the workload of a transistor is approximated with a statistical approach. Afterward, the aperiodic signal patterns are adapted to periodic patterns such that the AC NBTI model of [56] can be applied to estimate the threshold shift.
Different nodes within a digital circuit might have considerably different activity rates. In other words, not all nodes operate with high frequencies relative to the (high) clock frequency. Therefore, if the gate terminal of a PMOS transistor toggles rarely, the stress pattern applied to this node must be mapped to a much lower frequency than the clock frequency. At lower frequencies the induced ΔV_{th} is increased compared to higher frequencies [64, 65]. Extreme case would be similar to the DC stress pattern. It should be noted that for such low active nodes the recoverable component of ΔV_{th} is the crucial degradation mechanism and plays an important role in determining the transient performance degradation. This can be understood when short clock periods in practical applications (frequency in the range of several hundreds of MHz to GHz) compared to time constants of the NBTI recovery phenomena are used. Fig. 3.5 shows the ΔV_{th} degradation when applying stress/recovery sequences [66]. Besides DUF, the activity rate of each node is considered to take the recovery component of ΔV_{th} into account.
For the approximate AC NBTI model, the stress toggle rate at each PMOS gate terminal as well as the duty factor of stress is collected. This information about the activities of all transistors is gathered for a statistically significant number of inputs by the developed aging tool and is used to map aperiodic input patterns to periodic patterns as expected from the developed AC model. For large circuits the information about the activities of all transistors is gathered by applying a set of most likely patterns to the circuit. Therefore, the average duty factor DUF_{AV} and the average period T_{AV} over the applied stress pattern seen at a PMOS are calculated. The average duty factor is defined by

$$DUF_{AV} = \frac{NBTI_{stress}}{t_{AC}} \tag{3.13}$$

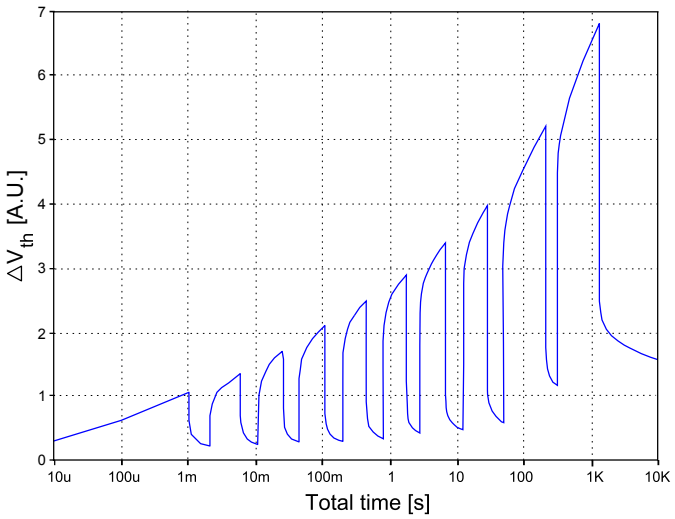


Figure 3.5.: NBTI degradation with an alternating stress-recovery pattern.

where $NBTI_{stress}$ denotes how often the stress condition was fulfilled for a transistor. $t_{AC} = n \cdot T_{clk}$ is the overall stress time where n is the number of applied clock cycles.
Although the frequency dependence vanishes at relatively moderate frequencies [15], it is crucial to consider both frequency dependency and the recoverable component of ΔV_{th} for nodes with low activity within the circuit. Therefore, it is additionally possible to differentiate between different stress shapes which result in the same duty factor. It is then assumed that one period consists of a stress phase followed by a relaxation phase. The average over all these periods results in

$$T_{AV} = \frac{n \cdot T_{clk}}{n_{stress-relax}} \tag{3.14}$$

in which $n_{stress-relax}$ specifies the number of stress relaxation cycles and can easily be calculated as half of the number of stress toggles on a certain gate terminal.
After collecting the statistics for every PMOS and performing the average of these values over the given number of clock cycles constant values are gained for the originally variable stress and relaxation times. With these constant values for DUF_{AV} and T_{AV} the model for a single transistor is applied and an individual constant threshold voltage shift is calculated for every PMOS. This shift is then added to the SPICE netlist as a voltage source in a sub-circuit by the developed aging tool. The clock frequency determines the length of recovery interval after stress for the ΔV_{th}. This enables a transient reliability assessment, preventing an underestimation of the guard-bands.

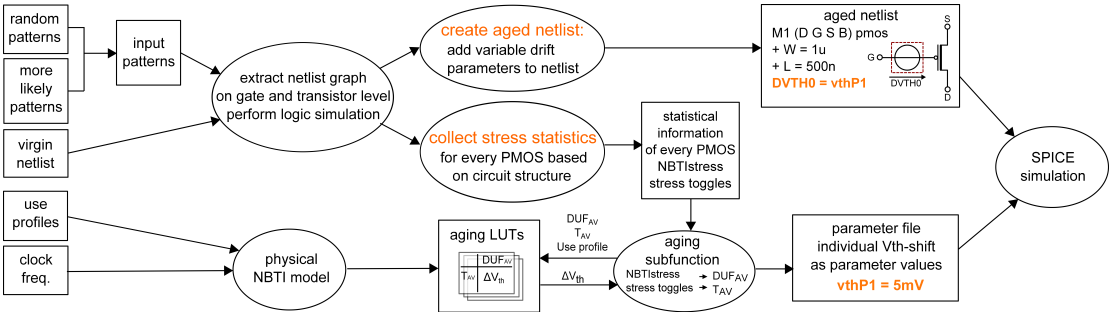


Figure 3.6.: Data flow diagram of the developed tool extrapolating aging for circuit lifetime

3.2.2. Circuit Level NBTI Simulation

Figure 3.6 shows the flow diagram of the developed NBTI aging simulation tool. The core of the developed aging tool is the modified analytic AC NBTI model. The aging tool reads in the structural information of a circuit from a SPICE netlist. Two inputs are provided including a fresh netlist and a set of input patterns. The set of input patterns is chosen either completely random for small modules or is a set of more likely ones for large circuits. Based on the HDL description of the standard library elements, logical functionality of each gate in the library is determined. The aging tool scans the netlist and establishes two graph structures on gate level and transistor level.

In the gate level graph the information about the connections between different logic gates and/or custom designed components is included. Thus, the gate level graph $G_{GL} = (V_{GL}, E_{GL})$ is formed, where vertex set $V_{GL} = \{v_1, v_2, .., v_m\}$ is the set of gates or custom designed components and edge set $E_{GL} = \{e_1, e_2, .., e_n\}$ is the set of gate level interconnects (nets). The gate level graph is directed as it is formed by directed edges (arcs). This is due to the fact that the CMOS combinatorial logic is unidirectional. The gate level graph also includes the set of primary input and output edges. Each vertex is linked to its corresponding library component through its Library property. It is also linked to its input and output edges through input and output edge properties. Each edge has several basic properties which are extracted from netlist. These include edge number, name, input and output vertices. Edge number and name are reference properties while input and output vertices refer to the vertices which are linked to this edge object. A valid property is required for the evaluation of graph edges and this valid property has the logical value of the edge after the evaluation of the gate level graph. Other properties such as number of toggles, number of zeroes and ones are used for evaluating statistical characteristics of each edge after evaluation of the circuit for a number of different input combinations. The NBTI aging property $NBTI_{Stress}$ is included to determine the characteristics of the edge for different stress conditions

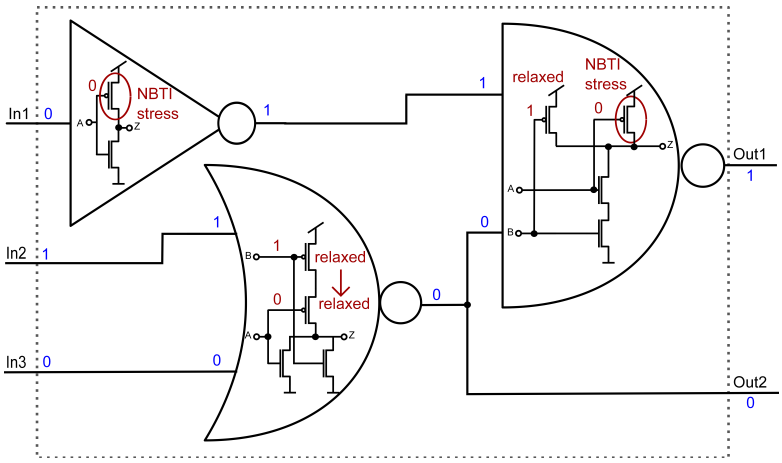


Figure 3.7.: A simple digital circuit as an example to demonstrate the logic propagation by the developed aging tool. On gate level logic values are propagated and on transistor level structure dependence of NBTI is regarded.

based on the circuit structure. Other aging properties can be easily included when the corresponding physical model is included into the tool.
In the gate level graph, logic values are propagated by utilizing the first depth search algorithm in the graph and digital signal simulation (see Fig. 3.7). If the component belongs to a standard library, logic simulation is accomplished by a functional VHDL description of the logic gates. Thus, the gate level description enables propagating signal values within the circuit which decreases the computational effort compared to transistor level simulation. If the component is custom designed the logic simulation is performed on transistor level.
The graph also considers sequential cells and is capable of digital signal simulation for synchronous designs with feed-backs.
The transistor level graph $G_{TL} = (V_{TL}, E_{TL})$ includes the transistor vertex set $V_{TL} = \{v_1, v_2, .., v_i\}$ as well as the transistor level interconnect edge set $E_{GL} = \{e_1, e_2, .., e_j\}$. The transistor level graph is undirected and the transistors within the graph are sorted from the voltage sources to the output. As an example the sorted transistors of a NOR gate are shown in Fig. 3.8. Based on the sorted transistor graph, switch mode simulation for custom designed circuits as well as stress evaluation for all components is performed.
The desired inputs are applied and propagated through the circuit where the statistics for NBTI stress for the gate terminal and the stress toggle rate for every PMOS transistor are gathered. The transistors that are identified such that they experience stress are indicated by ovals in Fig. 3.7.
Subsequently, the aged netlist will be prepared with parameters for the aging

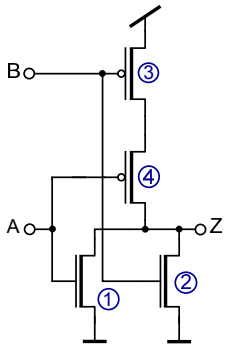


Figure 3.8.: Transistors in an exemplary 2-input NOR gate sorted to enable switch level simulation and BTI stress evaluation

sub-function. For every sub-circuit where a transistor experiences a ΔV_{th}, a new equivalent aged sub-circuit is written into the netlist. This sub-circuit comprises a voltage source at the gate terminal to incorporate the PMOS threshold voltage shift. The ΔV_{th} is then set as an individual parameter for each transistor at the corresponding voltage source. The parameter value of this voltage source is stored in a separate parameter file according to the parameter variable added in the netlist.
Next, a function with the NBTI model for a single transistor reads out the statistical information and calculates the average duty factor (DUF_{AV}) and average stress period (T_{AV}) for every PMOS. The corresponding individual ΔV_{th} for every PMOS is obtained from a look-up table (LUT) containing all possible threshold voltage shifts. These threshold voltage shifts are calculated and stored from the approximate AC NBTI model by a sweep over all possible T_{AV} and DUF_{AV} for a certain clock and use profile. For every transistor, the obtained statistical values for NBTI stress and toggle rates result in a specific DUF_{AV} and T_{AV} for which the closest corresponding ΔV_{th} is taken from the table. The table is dependent on T_{clk} and the defined use profile. The defined use profile includes the temperature, V_{stress} and the time span for which a circuit is aged. It is possible to create LUTs for different desired use profiles.
By considering the approximate NBTI AC-stress model, an evaluation over the desired lifetime is performed and the ΔV_{th} is written into a separate file containing all parameter values. Both the aged netlist and the parameter files are directly taken into a SPICE simulation, with the benefit that the statistics are gathered only once for a statistically reasonable amount of inputs. This is computationally feasible even for big circuits. For other aging scenarios with different use profiles only the parameter file is changed. Furthermore, the simulations retain their simulation speed due to the constant ΔV_{th} and hence no further large computational effort is added.

The developed tool enables reliability assessment of a circuit considering NBTI with a low computational effort and thereby provides a valuable means to predict the lifetime constraints of a circuit. Moreover, the developed aging tool is able to handle netlists with high level, gate and transistor level components. As an additional feature clock and power gating as used in state-of-the-art low power technologies are included.

3.2.3. Clock Gating

Clock distribution refers to distributing the clock signal from the phase locked loop (PLL) as the clock generator to the synchronous circuit elements. The clock distribution network can contribute up to 40% of the total power consumption of the digital circuit [67]. This is due to the high operating clock frequency and large load of the clock signal [68]. In addition to sufficient drive strength for the heavily loaded clock signal, the clock jitter and skew constraints need to be fulfilled during the required lifetime. The clock skew is defined as the maximum difference in the arrival times of the clock signal at different sequential elements which can interact with each other due to a connecting combinatorial path. The clock jitter is the temporal variation of clock edge with respect to the nominal undisturbed clock. The minimum clock period corresponding to the maximum frequency of the clock signal is then given by

$$T_{clk} >= T_{cq} + T_{pd} + T_{skew} + T_{jitter} + T_{setup} \tag{3.15}$$

where T_{cq} is the clock-to-Q delay of the flip-flop, T_{pd} is the propagation delay through the combinatorial path, T_{skew} is the clock skew, T_{jitter} is the clock jitter and T_{setup} is the setup time of the capturing flip-flop.
Therefore, the clock signal is one of the most critical signals in the digital circuit and a reliable clock distribution network has to satisfy the power limitations as well as the timing constraints (equation 3.15). Ignoring the temporal clock jitter, it can be seen from equation 3.15 that the clock skew has a direct impact on the maximum possible operating frequency of the digital circuit. Thus, decreasing the clock skew is a major design concern. Advanced clock trees often require a symmetric clock tree topology, e.g. H-tree. Moreover, complex signal routing algorithms are required to equalize the arrival times of the clock signal at all sequential elements. To meet the power constraints, the clock distribution networks also utilize clock gating techniques which deactivates unused parts of the clock tree. Clock gating components usually include a latch followed by an AND/OR gate [68]. The latch avoids glitches and premature ending of the clock signal. For a AND gate as the output stage the logic low is the controlling value, while for the OR gate the controlling logic high is used [69].

Thus, clock gating influences the extent of ΔV_{th} induced by NBTI for different paths in the clock tree. Moreover, clock gating increases the occurrence of nonuniform NBTI degradation. Asymmetrically aged clock buffers increase the clock skew and lead to the clock signals arriving at different times with bigger skews at the flip-flops. Without clock gating, the clock signals in all parts of the clock tree switch every cycle. That means that the PMOS transistors in clock buffers experience alternate AC stress and recovery phases of equal duration. On the contrary, the transistors that are a part of heavily gated clock buffers do not experience equally alternating stress. Thus, ΔV_{th} in gated clock buffers would be different from the remaining non-gated clock buffers. This nonuniform aging results in a large clock skew in the clock tree and might in turn result in timing violations.

3.3. Exemplary Use Profiles

3.3.1. Mobile Phone

The constraints for an exemplary mobile phone use case are defined as four years of operation at a maximum temperature of 85 °C and a 5% increased supply voltage compared to the enhanced value of $V_{DD} = 1.3$V. The specified product lifetime is considered as 4 years of continuous operation.

3.3.2. Automotive

For safety critical applications such as in automotive system errors due to performance failures must be avoided. Automotive products are designed to operate in harsh environments at high temperatures and for long lifetimes. Typical use cases are temperatures in the range of 85 °C-175 °C for a lifetime of 10-15 years [70]. Thus, aging as a temporal variation is enhanced significantly.
The conditions for an exemplary automotive use case for stress temperature and supply voltage are 125 °C and 105% of an enhanced $V_{DD} = 1.3$V, respectively. The specified product lifetime is considered as 10 years of continuous operation.

3.4. Test Circuits

3.4.1. Standard Library Components

As an example of the effect of threshold voltage shift on CMOS gates and to mimic dangerous scenarios within digital circuits, chains of standard core elements are characterized in a 65nm technology. Assume an inverter chain with low activity rate, i.e. the elements are in a steady state for long intervals. With high operating

frequencies, the relaxation time would be short and the recoverable part of NBTI has the dominant effect. Thus, for high activity nodes permanent NBTI and for low activity nodes both permanent and recoverable NBTI are considered. Here, the mobile use profile is used. Table 3.4.1 summarizes the delay increase after this use profile for slow corner with peak $Temp = 125\,°C$. The difference between NAND and NOR chain is due to the serial connection of NMOS or PMOS transistors, respectively.

	Permanent NBTI with periodic AC stress and frequency of 500MHz	Permanent and recoverable NBTI with short recovery time, shortly after steady state of 1h (similar to DC-stress with short recovery)
Inverter	2.3%	7.8%
NOR	2.6%	8.9%
NAND	2.2%	4.6%

Table 3.1.: Delay increase in standard core components (chain of gates)

3.4.2. 16 by 16 bit Multiplier

To demonstrate the applicability of the tool to complex circuits, first an arithmetic test circuit is evaluated in terms of reliability. The circuit under test is a 16 by 16-bit multiplier synthesized in an industrial design flow. The re-simulations of the aged netlists for the mobile use profile are performed at the nominal supply voltage of $V_{DD} = 1.2\text{V}$ at a moderate temperature of 27 °C for the following case studies including the standard instances and the multiplier. Figure 3.9 shows an example of the varied output diagram of three critical paths. Besides the delay increase, parameter shifts also affect the timing behavior of the circuit, e.g. resulting in deglitching or generation of new glitches. However, in Fig. 3.9 the final signal values do not result in errors, as they arrive to the flip-flops in time without violation of the flip-flop setup time. Fig. 3.10 shows the percentage propagation delay increase for the aged circuit in AC and DC mode with respect to the fresh circuit, which always does not exceed a % margin.

3.4.3. Secure Hash Algorithm

To demonstrate the ability of the developed aging analysis tool in detecting nonuniform aging, which is a result of structural dependence of NBTI, the secure hash algorithm (SHA-1) is implemented and synthesized. SHA-1 is a cryp-

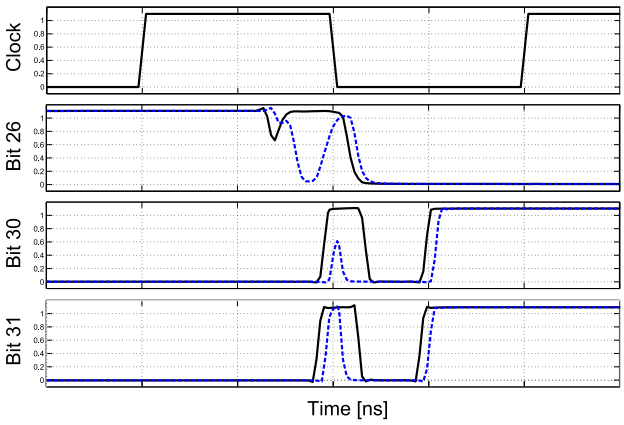


Figure 3.9.: Timing diagram for the outputs of the 3 critical paths belonging to a 16 bit booth multiplier circuit, when applying the worst case patterns. The fresh circuit shown in solid lines and the aged circuit in dashed lines.

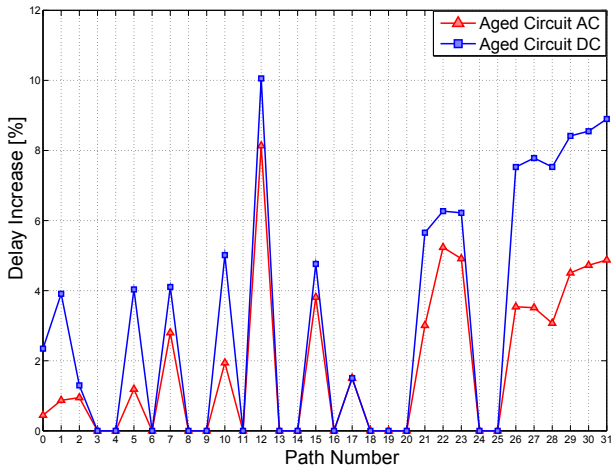


Figure 3.10.: Percentage of the delay increase for all paths of the 16 bit multiplier, worst case pattern simulation: AC-stress and DC-stress with respect to the fresh circuit. Performance simulation in nominal process, $V_{DD} = 1.20\text{V}$ and $Temp = 27\,^{\circ}\text{C}$

tographic algorithm used in the automotive applications. The synthesized SHA-1 circuit includes several high-level components. When investigating the SHA-1 circuit in terms of aging, nonuniform degrading paths were detected. An extreme nonuniform aging can reorder the critical paths and even result in a violation of the timing specifications if the safety margins were underestimated.
Nowadays, data bus systems connect nearly all electronic control units (ECU) in

vehicles. Since all these bus systems are coupled together, it is possible to have access to nearly all critical components within the car. To provide different safety relevant functions, several bus systems for communication are available. This means that, automotive communication networks have access to several safety related components of the vehicle such as brakes, airbags or the engine control. Moreover, cars equipped with driving assistant systems such as adaptive cruise control require extra functionality [71] depending on the underlying data networks [72].
Currently, modern cars are equipped with more than 80 ECUs [73], which are internally connected via serial buses and communicate using standard protocols. New car multimedia networks as well as wireless interfaces such as global system for mobile communications (GSM) or Bluetooth are introduced in cars which are coupled to the internal communication network. Since these multimedia networks are unsecured compared to the internal communication networks, there would be a possibility to externally corrupt internal critical functionality [74]. Moreover, recently the customer has the ability to send and receive data from the car via consumer electronics such as smart-phones or tablets. This also results in additional security risks [75]
Thus, to ensure confidential communication for the in-vehicle network, cryptographic algorithms such as SHA-1 is used [76]. The SHA-1 algorithm produces a 160-bit representation of a message, so called as digest, with variable length smaller than 2^{64}. Any change in the message results in a different hash value and the authentication fails to verify.
The SHA-1 algorithm includes two steps, the message padding and the hash calculation. SHA-1 processes only 512-bit blocks. Therefore, the message padding step is necessary to expand a message with variable length such that the resulting length is a multiple of 512. The hash calculation is based on a non-linear function f_i which operates on the 32-bit words m_0, m_1..., m_{15} contained in one 512-bit block. Figure 3.11 shows the calculation of the hash value in the SHA-1 algorithm.
Due to the structure of the algorithm, some parts of the circuit are in a steady state for a large number of clock cycles since some signals are constant for these times. Furthermore, the algorithm includes different functionality such that gates can experience a strongly varying workload and hence an asymmetric delay increase. The SHA-1 algorithm was implemented in VHDL and synthesized in a 65nm technology for a clock frequency of 500MHz (T_{clk} = 2ns). Clock gating is inherently introduced during the synthesis to reduce the operating power consumption. For registers which have a particularly low switching activity, clock gating is introduced, which also increases highly nonuniform degradation inside the circuit.

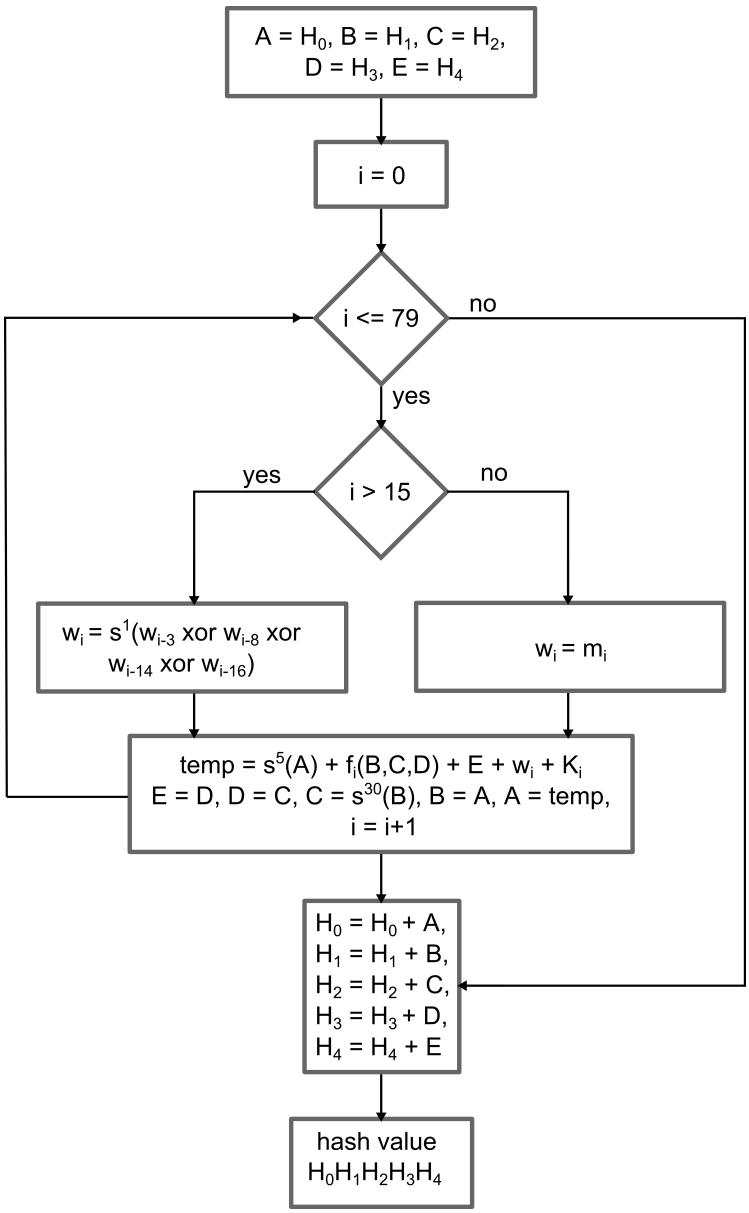


Figure 3.11.: Flow chart of calculating the hash value in the SHA-1 algorithm.

$H_0 = 67452301,\ H_1 = EFCDAB89,\ H_2 = 98BADCFE,$
$H_3 = 10325476,\ H_4 = C3D2E1F0$
$K_i = 5A827999,\ f_i(B,C,D) = (B \cdot C) + (\overline{B} \cdot D)$ for $0 <= i <= 19$
$K_i = 6ED9EBA1,\ f_i(B,C,D) = B \oplus C \oplus D$ for $20 <= i <= 39$
$K_i = 8F1BBCDC,\ f_i(B,C,D) = (B \cdot C) + (B \cdot D) + (C \cdot D)$ for $40 <= i <= 59$
$K_i = CA62C1D6,\ f_i(B,C,D) = B \oplus C \oplus D$ for $60 <= i <= 79$

Reliability Assessment of SHA-1

To evaluate the reliability of the synthesized SHA-1 circuit an exemplary automotive use profile was chosen. Supply voltage and temperature within the life span are considered 105% of an enhanced $V_{DD} = 1.3\text{V}$ and $Temp = 125\,^{\circ}\text{C}$, respectively. The circuit is considered to be under continuous operation for 10 years. The aging of the circuit is analyzed by the developed aging tool and the aged netlist is created. Afterward, the simulations of the fresh and aged netlist are executed at nominal and slow corners. The nominal corner is considered as nominal process, $V_{DD} = 1.2\text{V}$ and $Temp = 27\,^{\circ}\text{C}$. The slow corner is considered as slow process, $V_{DD} = 1.05\text{V}$ and $Temp = 175\,^{\circ}\text{C}$. Figure 3.12 shows the worst case

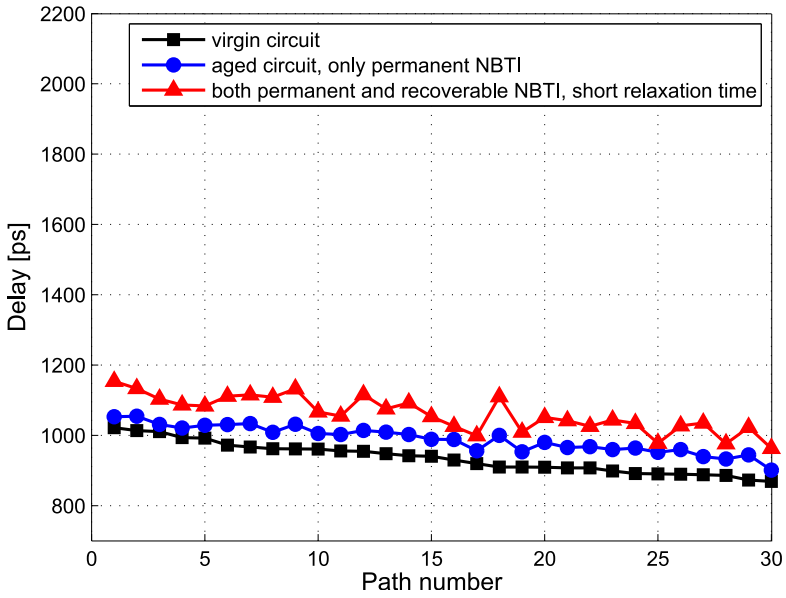


Figure 3.12.: Delays of the most critical paths for the fresh and the aged SHA-1 circuit, 10 years operation. Performance simulation in nominal process, $V_{DD} = 1.20\text{V}$ and $Temp = 125\,^{\circ}\text{C}$

delays of the 30 most critical paths with the highest delay for the fresh SHA-1 circuit. In this figure, the results of two aging regimes are depicted. In the first regime only the permanent component of NBTI is considered. In the second one, both permanent and recoverable parts of NBTI are present. As in the second regime the relaxation time is of the same order of magnitude as the clock period, due to the presence of the recoverable part of NBTI, larger ΔV_{th}s and thus bigger performance change is observed. As shown in Fig. 3.12, assigning different recoverable NBTI shifts for gates with very different workloads results in nonuniform aging. However, in the nominal corner, aging does not result in critical operation. This is due to the fact that an increased delay of the combinatorial path does not

violate the setup time of the capturing flip-flop. Therefore, to take into account PVT variations, simulations are performed for slow corner (Fig. 3.13). In the slow corner, a voltage drop of 150mV and a short interval with a peak temperature of $Temp = 175\,°C$ are considered. Considering the corner case, longer delays occur due to slower devices and lower V_{DD}. Thus, a large performance degradation is observed. Considering the setup time of an aged commonly used scan flip-flop at the slow corner, ($\approx 150 - 200\,ps$) a setup time violation occurs resulting in errors at the output of the flip-flop. However, assuming a full time circuit operation is

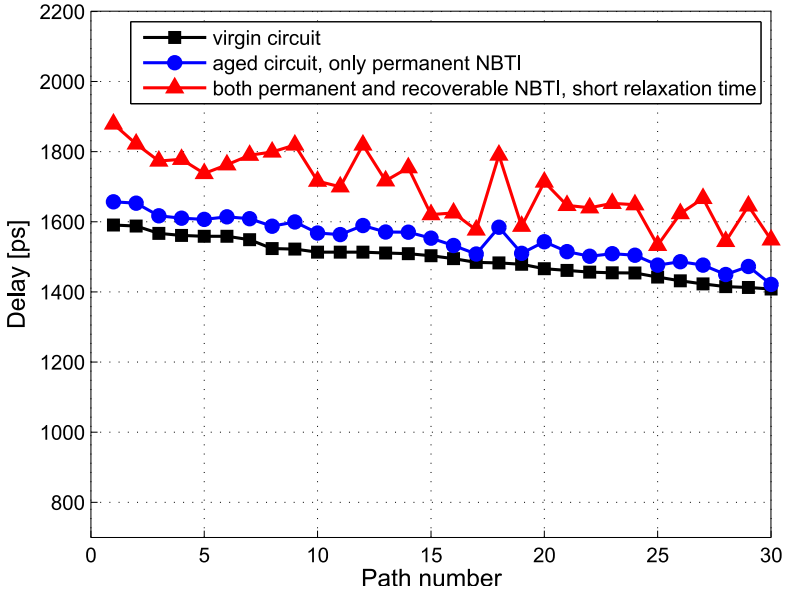


Figure 3.13.: Delays of the most critical paths for the fresh and the aged SHA-1 circuit, 10 years operation. Performance simulation in slow process, $V_{DD} = 1.05V$ and $Temp = 175\,°C$

not realistic. Thus, another regime needs to be evaluated to take care of the permanent NBTI over the life span of the circuit and both recoverable and permanent effects under short operation times after a large recovery time (Fig. 3.14). Here, two operating times (60s and 3600s) are considered. The delay degradations are smaller than Fig. 3.13 but yet considerably larger than only permanent NBTI. The remaining timing slack within a clock period is very small. Thus, the setup-times of the capturing flip-flops might be violated and a reliable operation is not provided. Table 3.2) summarizes the resulting data for permanent NBTI ($NBTI_P$), recoverable and permanent NBTI ($NBTI_{P,R}$) after over 10 years. Moreover, it summarizes the simulation results for permanent NBTI ($NBTI_P$), recoverable and permanent NBTI ($NBTI_{P,R}$) after over 10 years and operation times of 60s and 3600s, respectively.
Thus, with the developed aging analysis tool, an accurate evaluation of the aging

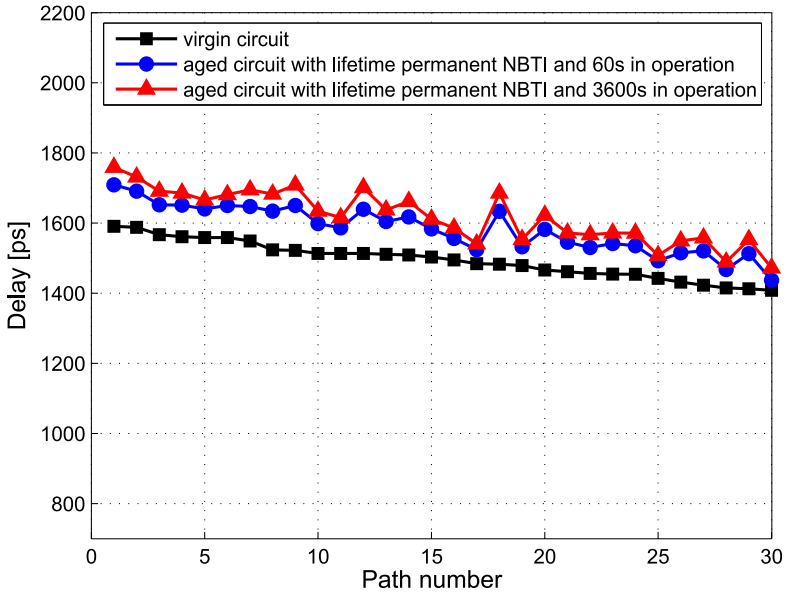


Figure 3.14.: Delays of the most critical paths for the fresh and the aged SHA-1 circuit, permanent NBTI ($NBTI_P$) over 10 years, recoverable and permanent NBTI ($NBTI_{P,R}$) for operation times of 60s and 3600s. Performance simulation in slow process, $V_{DD} = 1.05$V and $Temp = 175\,°C$

Aging regime	Delay increase
fresh circuit slow corner	-
$NBTI_P$ 10 years continuous operation	6.9%
$NBTI_{P,R}$ 10 years continuous operation	20.7%
$NBTI_P$ in 10 years continuous operation and $NBTI_{P,R}$ in 60s	10.2%
$NBTI_P$ 10 years continuous operation and $NBTI_{P,R}$ in 3600s	13.7%

Table 3.2.: Performance degradation due to NBTI for SHA-1 circuit, permanent NBTI ($NBTI_P$) over 10 years, recoverable and permanent NBTI ($NBTI_{P,R}$) for operation times of 60s and 3600s

of the devices and the impact of device aging on circuit performance over lifetime is possible. The recoverable NBTI also needs to be evaluated to identify the potential reliability threats. In the developed aging tool, by using a physical NBTI model, individual threshold voltage drifts were determined for each transistor. By updating the physical NBTI model for new technologies it is possible to easily update the tool for scaled circuits. This means that integrating new device models

for new technologies is highly efficient. Moreover, by combining gate level and transistor level approaches, the tool gathers activity information on every node within the circuit on gate level. Afterward, the tool uses the physical NBTI model on transistor level to determine the threshold voltage drifts. Therefore, the tool is more accurate than the gate level approaches. Moreover, the developed aging tool is capable of handling both custom designed and complex synthesized digital circuits with lower computational effort compared to the purely transistor level approaches.
The developed aging tool can be used to determine the optimal placement of in situ monitors within the circuit. By using these monitors, reliability can be diagnosed and ensured during the lifetime (see chapter 4). The re-simulated aged netlist can be used to find new critical paths which arise due to nonuniform aging and monitors can be placed at these additional paths. By adding additional monitors at paths with nodes that show a particularly high NBTI sensitivity, errors due to nonuniform aging can be detected even without a re-simulation of the aged netlist. Moreover, weak spots within the circuit can be identified and modified during the design phase to ensure specific reliability requirements.

3.5. Summary

A novel aging tool was developed which is applicable on circuit level and is able to extrapolate the aging induced ΔV_{th} over the lifetime of a digital circuit. The tool considers state-of-the-art design techniques like clock gating. Besides the NBTI permanent parameter shift, the recovery of NBTI is considered. Thus, the tool is able to detect short term performance degradation due to the recoverable component of NBTI. As test cases, the aging tool is used to analyze the reliability of several circuits including a safety critical circuit. When analyzing this circuit with the developed tool, it is possible to detect asymmetrical aging. Therefore, the developed aging tool provides a useful means to encounter generous guardbands, which determined by a worst-case approach, while preventing an early circuit failure. Moreover, the developed aging tool can be used to determine the weak spots of the circuit to be used for ensuring the reliability during the lifetime by in-situ monitors or component replacements.

4. Reliability Management by in situ Monitoring

As aging results in performance reduction of digital circuits, timing properties of the monitored circuit can be used as an indicator for the degradation level and reliability status of the circuit. Other defect cases such as inline resistive faults are also detected by monitoring the timing of the circuits. The reason is that finite resistance interconnects and intra-gate opens introduce additional delays in the circuit. Therefore, paths exposed to risk of malfunction can be detected by analyzing the timing properties [77].
To resolve the problems of global speed monitors using replica paths, which do not represent the real circuit to be monitored, in the reliability monitoring developed in this work in situ delay monitors are exploited [50, 7, 6]. By observing the timing properties of the monitored circuit in situ, within die variations of process, voltage, temperature and aging (PVTA) are monitored and an accurate assessment of the reliability status of the circuit is possible.
This chapter provides a system level perspective for the monitoring methodology. In order to monitor the timing of the circuit two approaches are possible: monitoring the circuit during functional operation and/or during test sequences, which are regarded in this chapter. Moreover, this chapter discusses an approach to consider different parts of the circuit with different reliability criteria.

4.1. Offline Monitoring

Offline monitoring refers to monitoring the circuit during dedicated test sequences. During the assigned time intervals, the circuit enters the test mode in which required test patterns are applied to the circuit and the resulting test outputs are collected and observed.

4.1.1. Design for Testability

Testing of digital circuits at the beginning of the lifetime is a necessity due to the possibility of an imperfect manufacturing process. To have a good testability, design for testability (DFT) methods are introduced. By applying DFT methods the costs of test development as well as the testing execution time are reduced.

An example of the chip level DFT technique is Built-in Self-Test (BIST), in which test patterns are generated and applied to the circuit under test (CUT) by on-chip hardware [78].
A systematic DFT approach is the scan based BIST, which converts the sequential circuit into a combinatorial design. This combinatorial design is then tested. A scan based design can be operated in functional (normal operation) mode or test mode activated by the Scan-Enable signal. When the Scan-Enable is disabled, the scan cells operate as normal D-flip-flops. When the Scan-Enable is activated all the scan cells are connected and form a shift register. Two different methods are introduced for scan based BIST, test per scan and test per clock [79]. In test per clock scan based BIST, a test vector is applied and test responses are captured at every clock cycle. In test per scan BIST, the test vector is shifted into the scan chains and a functional cycle captures test responses after shift cycles.

4.1.2. Transition Delay Fault Testing

Different scan based delay fault tests have been introduced to diagnose transition faults [80]. The delay fault testing through scan based design provides a means to diagnose the propagation delay of the paths under test at the early time of the life of the chips, i.e. after production. Delay fault testing is essential to ensure reliable and high quality products for safety critical applications.
In the basic transition delay fault (TDF) model [81] rising/falling transitions at each node are tested. The two-pattern test $\langle v_1, v_2 \rangle$ generates a rising/falling transition at the certain node. The initialization vector v_1 is applied to the circuit and all signals retain their initial value. Afterward, the propagation vector v_2 is applied to the circuit and the outputs are sampled at-speed [82]. For the node under test v_2 is a stuck-at $0/1$. The delay fault is detected if it exceeds the clock period. When applying the two pattern test $\langle v_1, v_2 \rangle$ into the serial structure of scan chains, vector v_2 is either a one bit shift of $v1$ (launch-on-shift), or the circuit's response to v1 (launch-on-capture) [83, 84]. Figure. 4.1 shows clocking of the launch-on-shift and launch-on-capture scan tests.
In the launch-on-shift scan test, the transition is launched in the last clock cycle of the shift operation and the transition is captured by the system clock pulse. The scan enable switches at speed, which requires special attention to the routing of the scan enable signal [85]. As the launch path is the scan path, the test is more controllable. In the launch-on-capture approach, the transition is launched from the functional path. Thus, both launch and capture edge receive the data on the system side of the flip-flop and not the scan input. In other words, both launch and capture vectors are controlled by system clock pulses. The scan enable does not need to switch at-speed and is not critical for the timing. However, as the launch path is the functional path, the launch-on-capture approach is less controllable than the launch-on-shift approach.

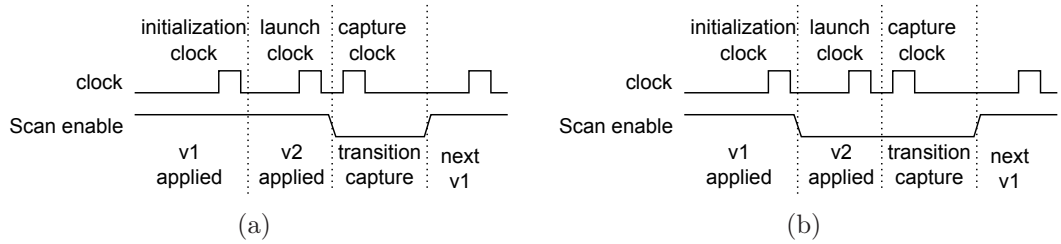


Figure 4.1.: Scan testing by a) launch-on-shift and b) launch-on-capture method

TDF tests are similar to stuck-at tests and the test coverage approaches stuck at fault coverage. Here, launch-on-shift shows a better coverage than launch-on-capture, but with the disadvantage that at-speed switching of scan enable signal is required. However, this is normally not the case in most designs with multiplexer based scan flip-flops [83].
It should be noted that in our lifetime monitoring method only a small percentage of the paths are equipped with monitors (less than 10%). Thus, for lifetime delay monitoring it is sufficient to cover only the transition test of the paths equipped with monitors. For the monitoring system in this work launch-on-capture method is assumed to relax the routing requirements of the Scan-Enable signal. However, the application of scan test routine is out of the scope of this work.

4.1.3. Scan Based Design Equipped with Monitors

The aging mechanisms affect the circuit parameters gradually. Thus, it is highly advantageous to be able to monitor the circuit during its lifetime to be able to assess its reliability status. This section proposes a method to utilize the existing scan chains for lifetime monitoring purpose.
Aging monitors are inserted within the scan path, as shown in Fig. 4.2. During the test sequences, aging critical test vectors are loaded into the scan chain using the Scan-In (Test-In) signal. These test vectors need to stimulate the paths equipped with monitors. After applying specified critical test inputs the timing information of the circuit is captured by the inserted monitors [50, 86]. The extracted timing information is related to the transition time of the output of the combinatorial logic, i.e. the delay of the path under test. In this approach, monitors are only enabled during the test sequences, e.g. by a Monitor-Enable signal. Monitors can be designed in a way that by disabling the monitors the dominant stress condition is removed. Since the testing interval can be chosen short compared to the normal operation of the circuit, aging of the monitors is negligible.
It is possible to store the information extracted by the monitoring system on an embedded non-volatile memory. This enables comparison to the worst case admissible slack also stored in the memory and thus decision making regarding

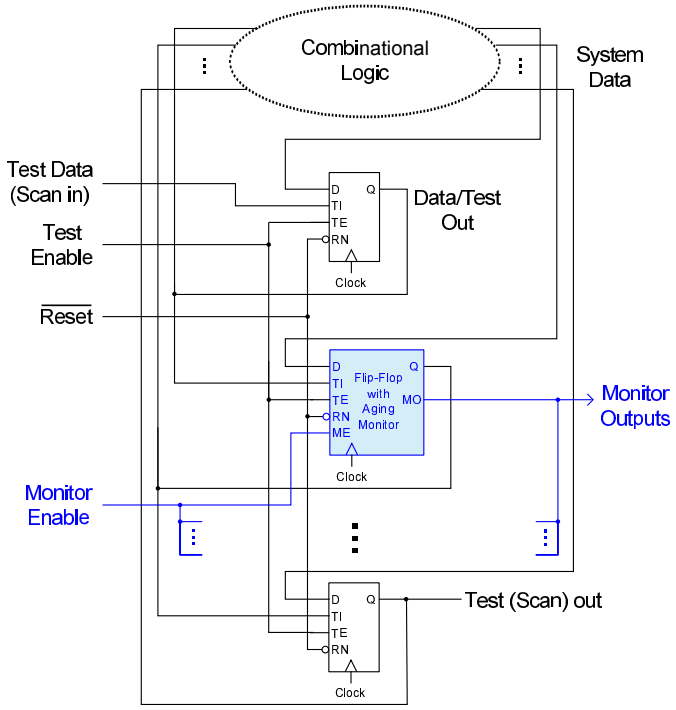


Figure 4.2.: Scan based design equipped with aging monitors

the reliability status of the circuit.

4.1.4. Efficient Monitor Placement

To minimize the power and area overhead of the monitoring system while providing sufficiently accurate data, the in-situ timing monitors are placed at optimal locations. To place the aging monitors, one approach would be to replace regular flip-flops at the end of the most critical paths by aging monitors. The number of critical paths used in this approach could be determined by choosing a timing margin for performance degradation over lifetime. As aging is not considered in the selection process of the paths, typically too many monitors are included, leading to unnecessary power and area consumption.
Another approach to obtain timing information regarding the degradation of the circuit, is to place the aging monitor at the end of only one critical path. As different circuit paths degrade by different extents, critical paths may change during lifetime and this approach may lead to deficient and incomplete aging monitoring. Hence, to have a more reliable monitoring scheme, aging monitors are inserted at the end of critical paths suffering from most degradation during lifetime, resulting in the most reduced timing slack. For big circuits the critical paths can be identified by commercial tools such as PrimeTime. Alternatively, such paths can

be identified during design phase by aging aware static timing analysis [35].
In smaller circuits, these critical paths can be distinguished more accurately by the developed aging simulation tool presented in Chapter 3. Therefore, safety critical parts are identified to optimally place monitors in the circuit. By comparing the extracted timing information to the worst case admissible slack, the reliability status of the circuit is determined and required countermeasures are carried out.

4.1.5. Test Pattern Generation

Critical test patterns need to be identified to stimulate selected paths equipped with monitors. Therefore, the selected paths are provided to an automatic test pattern generator (ATPG) software in order to obtain aging-aware test patterns. The generated test patterns are then stored on a non-volatile memory. These test patterns are applied to the circuit by the existing scan paths equipped with monitors.

4.2. Online Monitoring

Compared to the offline monitoring method [50, 87, 6] and BIST approaches [88, 89], When monitoring the circuit under functional operation, no interference with the operation of the circuit is required. Therefore, idle times for testing are avoided. Moreover, power dissipation of the monitoring system is reduced. Another major advantage is that for aging monitoring the very same operating parameters (supply voltage, clock frequency and temperature) as the functional circuit are present. The online monitoring approach can be activated from time to time to perform tests on the functional circuit.
To minimize the aging of the monitor circuitry, monitors have a disabled mode in which the dominant stress condition is removed. Similar to the offline monitoring method, since the testing intervals are chosen short compared to the normal operation of the circuit, aging of the monitor circuitry is negligible. However, it is possible that a circuit does not experience critical transitions during the monitoring sequence and thus resulting timing information might not show the actual criticality of the reliability status. Thus, monitors must be located in way to achieve the most accurate data regarding the system reliability level. Therefore, the placement of monitors is critical in this approach and requires high attention.

4.2.1. Efficient Monitor Placement

Similar to the offline monitoring approach, monitors should provide accurate information regarding the degradation of the circuit, while having small penalty in terms of area and power. Therefore, it is only feasible to equip a limited number of paths with the online monitors and it is necessary to find a solution for efficient placement of the monitors. In other words, monitors should be placed in locations which result in extraction of the most accurate data regarding the current reliability status of the circuit under test.
To optimally place the online monitors the simplest solution would be to select the most critical paths during the lifetime. It should be noted that individual paths are stimulated with a divergent frequency of transitions. This leads to different stress scenarios. A path with a low transition frequency or a steady state is more prone to timing violation by recoverable part of NBTI. The reason is that even though the transition probability is low, the delay of the path needs to satisfy the setup time constraint of the capturing flip flops for relatively high frequency clock. On the other hand, a path with a high activity rate might experience more transition at the capturing flip flop. This makes the path a suitable place for the monitor, as there is a higher probability for the appearance of a transition. In this section the solution for selecting the critical paths to be equipped with online monitors is introduced. The flow diagram of the monitor placement algorithm is shown in Fig. 4.3.
The SPICE netlist in the flow diagram is the synthesized netlist generated from the HDL circuit description in the semi-custom design flow. In the first step of the digital design flow, the register-transfer level (RTL) code (HDL circuit description) of the circuit is implemented. The RTL level code is synthesized into the circuit netlist by the semi-custom design flow. The synthesized netlist is on gate level which lists all the components of the circuit comprising of certain standard library elements. Afterward, a representative workload is used to perform the system logic profiling. Thus, a set of input patterns are applied to the circuit under test. The logic values are propagated through the circuit. Place and route is an important aspect at the generation of the post layout netlist in which the individual gates are ordered concretely. The aim here is to minimize the area consumption whereas the chip timing has to meet the performance requirements. The parasitic delays of the wires play an important role for the overall path delays. Thus, connected gates are placed with the minimum distance possible. In many cases it is not possible to minimize the wire length for all paths. Here, signals with higher switching activity are equipped with shorter wires.
To select the paths to be equipped with the monitors, first a high number of the potential paths in the circuit in the nominal corner are considered. Afterward, these potential paths are examined for the impact of process, voltage and temperature and also aging variability during the lifetime. Moreover, this work proposes

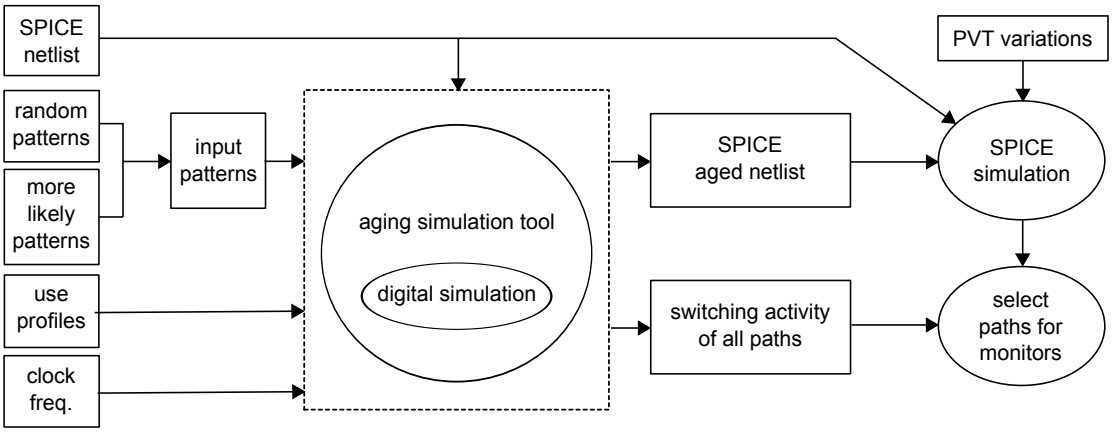


Figure 4.3.: In situ monitor placement flow diagram for the online monitoring method

that for an online monitoring system the transition probability, i.e. switching activity (SA), plays an important role in selection of paths to be equipped with the monitors.

Potential Paths to be Monitored

To demonstrate and evaluate the monitor placement algorithm a 16 by 16 bit multiplier circuit is chosen. To determine the potential paths to be monitored, a SPICE simulation with a significant number of random input patterns (1000) is performed.
The maximum delay within the circuit is denoted as $t_{d,\,max}$. However, the probability of occurrence of such a critical timing transition in the most critical path is very low (here 0.1%), as can be seen in Fig. 4.4. Thus, observing such a transition by the in situ monitors might be very time consuming. It might be that the maximum delays of the other paths occur more frequently and thus can be monitored earlier. Moreover, by only monitoring the most critical path, it is possible that the other paths produce errors, even though a critical transition is not yet detected in the most critical path. Therefore, in addition to the most critical path, paths with a high probability for critical delays are also regarded. By this approach, more transitions within the critical delays are observed by the in situ monitors and timing criticality of the circuit is detected faster.
In the first step, to define the critical transitions, a window within the clock period ΔT is defined and a transition is considered as critical if $t_{d,\,crt} >= (1 - \Delta T/t_{d,\,max}) \cdot t_{d,\,max}$. As opposed to the pre-error detection window approach [9], the ΔT is not a fixed value and does not correspond to the clock period, but depends on how close the delay of a path is to the maximum delay occurring

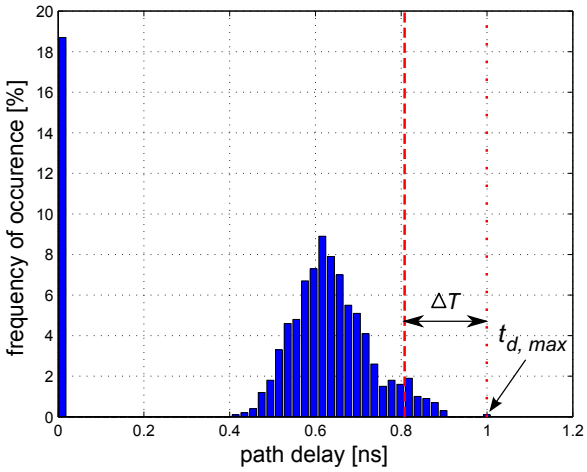


Figure 4.4.: Frequency for the occurrence of different delays of a critical path and the threshold at the beginning of ΔT. The maximum delay within the circuit is denoted as $t_{d,\,max}$.

in the circuit. This is similar to the path-based reduction step in [35], in which potential critical paths are identified for the aging aware static timing analysis (STA). However, here in contrast with [35], ΔT as a guard-band is introduced and $t_{d,\,max}$ is the maximum delay for the fresh circuit. The effect of an aging induced performance degradation is considered in the next steps. This is due to the fact that the potential paths are not only identified in terms of timing criticality during the lifetime, but also the observability of the occurrence of a critical transition is crucial, i.e. what would be the critical transition rate at the end of such paths. For a very small ΔT compared to the $t_{d,\,max}$, the probability of the occurrence of a critical transition would be low, and the information extracted by the monitors might not show the actual criticality of the timing properties of the circuit under test. Fig. 4.4 shows the threshold for identifying a critical transition when $\Delta T/t_{d,\,max}$ is chosen as 0.2. All transitions on the right side of the threshold (dashed line) are defined as transitions with critical delays ($t_{d,\,crt}$).
Paths which exhibit critical transitions are nominated as the potential paths (PP). For these defined potential paths, the average critical delays are calculated. To simplify the analysis, values are normalized.

Path Switching Activity

To efficiently select the paths equipped with the monitors, the node switching activity plays an important role. As mentioned before, paths with a low switching activity can be prone to the recoverable part of NBTI. On the other hand, paths with high activity rates are well suited to detect timing criticality by an online

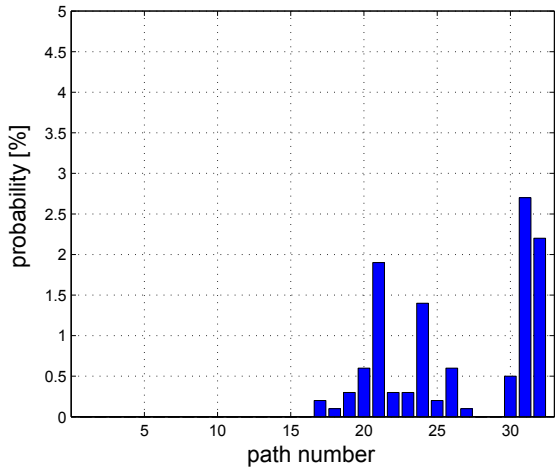


Figure 4.5.: Probability (%) for the occurrence of critical delays of the potential paths, $t_{d,\,crt} >= (1 - \Delta T/t_{d,\,max}) \cdot t_{d,\,max}$ in which $\Delta T/t_{d,\,max} = 0.2$

monitoring system. The reason is that the probability of a data transition at the end of such paths is higher than at the others. The goal in the developed algorithm is to find the spots in which the timing criticality can be detected with higher probability. Here, only the input patterns resulting in critical delays are of interest. Consequently, the probability for a transition at the outputs of the paths with critical delays are determined and depicted in Fig. 4.5.

PVT Variations

To consider the effect of process, voltage and temperature (PVT) variations simulations in different corners are performed. Thereby, only the transitions with a delay longer than $(1 - \Delta T/t_{d,\,max}) \cdot t_{d,\,ref,\,n}$ are treated, in which $t_{d,\,ref,\,n}$ is calculated for each specific corner case n. Here, $t_{d,\,ref,\,n}$ is the largest delay of all paths in the corner case n. The normalized average values are then calculated for each potential path.

Aging

For the SPICE netlist, the reliability assessment tool presented in Chapter 3 predicts the reliability status of the circuit during the lifetime. Therefore, the sensitivity of the potential paths to permanent and recoverable NBTI is evaluated. The results of the aging evaluation as well as the corner analysis and the statistics regarding the switching activities are considered in the developed algorithm for choosing the paths to be equipped with the monitors, as explained in the following.

Path Selection

To select from the potential paths efficiently and reliably, the criteria regarding the PVTA variation and the switching activity are considered. Based on a weighting function the most suitable paths are identified.
The weighting function for the insertion of the monitors is defined as

$$w_{m,\,PP} = \sum i \cdot t_{d,\,norm,\,i} \tag{4.1}$$

in which $i \in \{P, V, T, A, SA, D\}$ is the coefficient factor for considering each design criteria (process, voltage, temperature, aging, switching activity and mean of delays, respectively). $t_{d,\,norm,\,i}$ shows the normalized delays for each criteria. Here, $t_{d,\,norm,\,D}$ shows the normalized mean of the delays over all transitions.
Based on the number of the monitors to be placed within the circuit, the paths with the maximum values of the weighting function, $w_{m,\,PP}$, are chosen. For our simulations $\Delta T/t_{d,\,max} = 0.15$, the factors $P = V = T = SA = D = 1$ and $A = 2$ the results for the potential paths can be seen in Table 4.1. Finally, the 4 most suitable locations for placement of monitors by the largest values of $w_{m,\,PP}$ are identified as paths 22, 31, 32 and 24.

Path	22	31	32	24	20	21	16	23	30
$w_{m,\,PP}$	5.75	5.42	5.01	4.98	4.56	4.54	4.33	4.28	4.28

Path	17	25	26	18	19	27	28	29
$w_{m,\,PP}$	4.28	4.10	4.02	3.89	3.88	3.60	3.56	3.04

Table 4.1.: Weighting function for insertion of a monitor at the end of a potential path ($w_{m,\,PP}$)

4.3. Section Based Design

For a complex system it is advantageous to divide the system into several sections with different criteria for reliability (e.g. acceptable error rate). Several sections are formed which are a set of outputs of combinatorial paths. Consider a scenario in which for a specific section, errors are not acceptable, but for another section a certain error rate is tolerable. For instance, for a system comprising finite state machines and several data processing units, the finite state machines have higher reliability requirements. Another scenario could be that a section is under extreme stress conditions, resulting in more rapid degradation than for the other sections. Consequently, future errors might appear earlier for this section than other sections.
Considering different criteria for reliability, a customizable monitoring system can

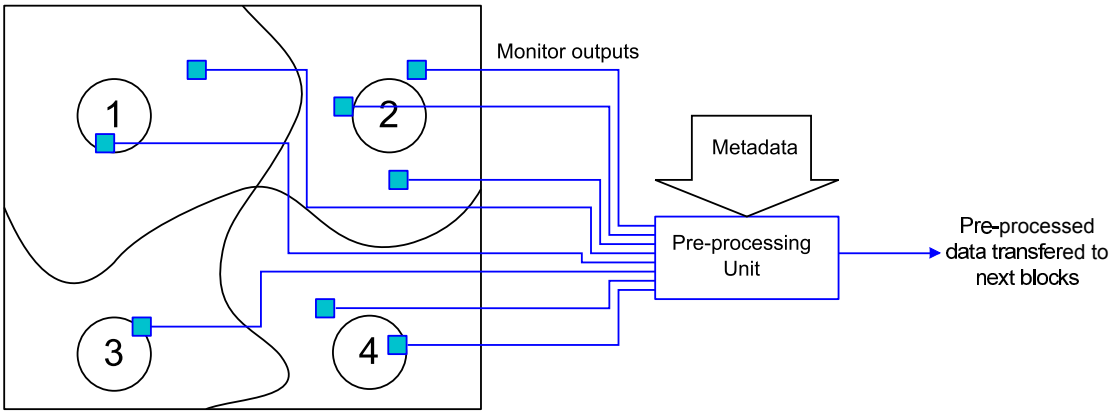


Figure 4.6.: Section based monitoring system with different reliability criteria for each section

be developed in which separate sections are considered in the monitoring process. These criteria form the metadata for each section and are used to check the correct operation of the system. Metadata can contain information such as maximum allowed delay (minimum slack), maximum error rate etc. In case of the first scenario, with different acceptable error rates, a criterion is defined for each section. This metadata defines the acceptable operation of the entire system, which could be for example the acceptable error rate.
The pre-processing unit can be programmed and designed to take into account metadata for different sections. An example of such pre-processing is pulse tuning for slack measurements.
Taking into account the defined metadata to design the pre-processing unit can be either in a dynamic or static manner. In case of dynamic tuning, representation of metadata as an input word is applied. This is used for tuning of pre-processing unit. In case of static design, metadata is applied to the pre-processing unit during the design phase, and therefore the pre-processing unit is customized for a certain circuit. Fig. 4.6 shows the section based monitoring system, comprising in-situ monitors and a pre-processing unit for manipulation of monitoring data for different sections with different reliability criteria.

4.4. Summary

In this chapter two different monitoring approaches were discussed: monitoring during the normal operation and monitoring during the test sequences. The required criteria regarding the monitor placement in both approaches were discussed. A novel monitor placement approach for inserting the monitors within the circuit

in an online monitoring scheme was proposed. Moreover, the criteria to take into account the different reliability demands for different parts of the section were discussed. Next chapter discusses the developed necessary circuits for the monitoring approaches.

5. Required Circuitry for in situ Reliability Monitoring

The in situ delay monitors extract information regarding the circuit level timing properties. Timing information extracted by the monitors is converted to the digital domain and transferred to higher layers of abstraction to diagnose the level of degradation and thus reliability of the system. Therefore, potential reliability threats are detected and the necessary countermeasures are performed. By taking such countermeasures, the lifetime of the circuit is prolonged and possible system failures are avoided. In addition, in case of predicting a non-reparable failure in the near future, a maintenance signal can be generated. The rest of this chapter discusses the required circuitry for precise monitoring of timing properties.

5.1. Design of in situ Timing Monitors

In this section different approaches for the implementation of the in situ monitors are discussed.

5.1.1. One Bit Monitors

In the pre-error approach [8, 90, 9, 22], in-situ delay monitors with the ability to distinguish between relaxed and critical operation of the circuit are used as timing monitors. These critical transitions, called pre-errors, indicate a reduced timing slack and thus performance degradation. In the beginning [22, 8], the pre-error detection approach was used for adaptive voltage scaling (AVS). However, the monitors used in the pre-error AVS can be modified for monitoring the performance degradation due to device aging.
During a certain time interval before the clock rising edge, called the pre-error detection window, data transitions result in a pre-error signal. Therefore, the pre-error flip-flop represents a one bit time to digital converter (TDC). The pre-error detection window length can be adjusted to meet different reliability requirements. Either the clock duty cycle or a delay element can be utilized to implement the pre-error detection window.
Different approaches for pre-error monitors have been evaluated [91] in terms of

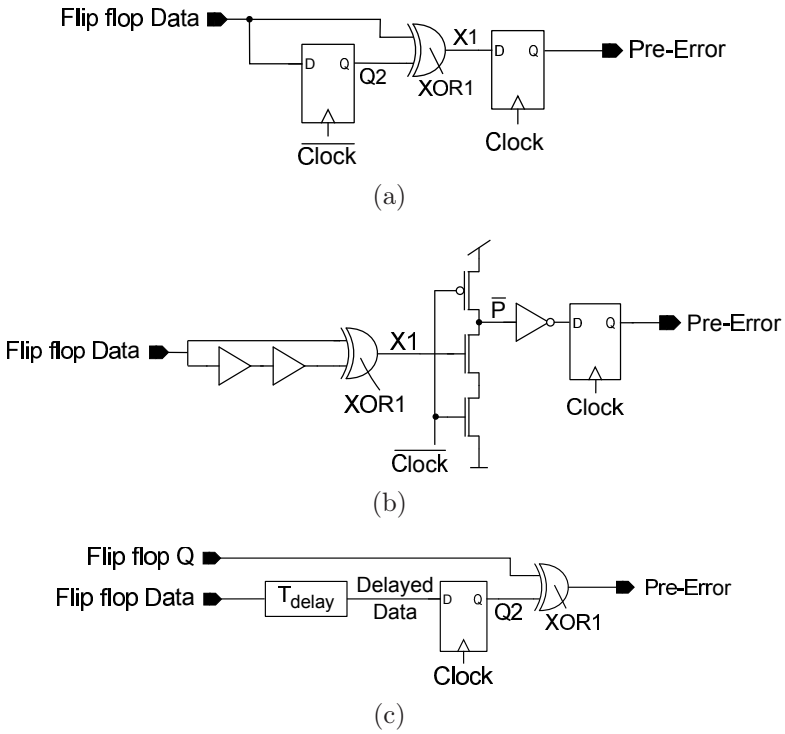
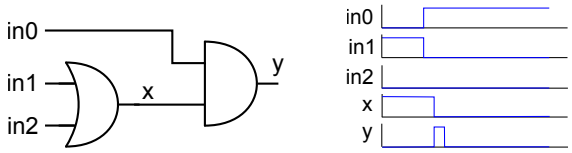


Figure 5.1.: Different pre-error monitors: a) duty cycle based static pre-error monitor b) duty cycle based dynamic pre-error monitor c) delay element based pre-error monitor

accuracy and robustness. Fig. 5.1 shows different approaches for the pre-error monitors.

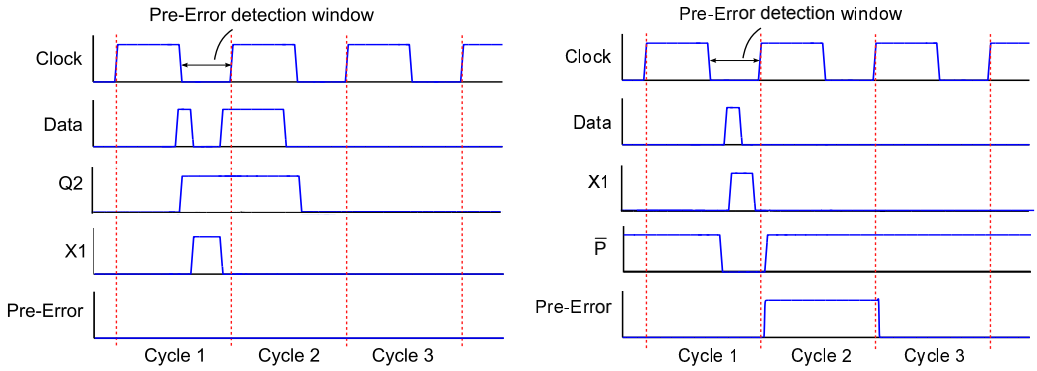
Glitches

A glitch is an undesired transition that occurs before the intended stable value is reached in digital CMOS circuits. A glitch occurs in digital circuits when the differential delay at the inputs of a logic gate is greater than the inertial delay [92]. Fig. 5.2 shows an example of the occurrence of a glitch in a digital circuit.
Fig. 5.3a shows the problems occurring for a static approach for the design of the pre-error flip-flop. Occurrence of a glitch during the detection window of the static pre-error flip-flop might not be observed by the pre-error monitor even though it might result in timing errors.
Fig. 5.3b shows the occurrence of a glitch for a dynamic approach for the design of the pre-error flip-flop. As the data should be stable before the triggering edge of the clock signal, the glitch might result in timing errors and must be assigned as a pre-error. However, it should be noted that in the dynamic pre-error approach the occurrence of the first transition of the data signal triggers the pre-error detection,



(a) (b)

Figure 5.2.: An example of the occurrence of a glitch in a digital circuit. Inputs (in0, in1 and in2) are assumed skew free.



(a) (b)

Figure 5.3.: Two scenarios of the occurrence of a glitch for a) static and b) dynamic pre-error flip-flop approaches.

even though the second transition is more critical in timing. However, it should be noted that the aging mechanisms such as NBTI have a slow gradual nature. Thus, when the performance is decreased to a degree that a glitch can become critical the pre-error monitor is able to identify the problem (the second transition in Fig. 5.3b).

Aging Resistance

A global signal can be introduced as Monitor Enable which is fed to the in-situ aging monitors to select between either monitor mode or normal operation of the circuit without the monitors [6, 23]. By disabling the Monitor Enable, in-situ monitors are not under stress conditions and thus do no degrade. Moreover, monitors enter the recovery phase in case of retrievable aging effects, e.g. NBTI. Switching off monitors during normal operation also minimizes their power overhead. Figure 5.4 shows the delay based pre-error monitor with Monitor Enable signal for aging resistance.
The Monitor Enable signal removes the stress condition when disabling the monitors. For this purpose, it is necessary to identify the dominant aging mechanism

for the monitor structure in the specific technology. NBTI is considered as a dominant aging mechanism in 65nm and 40nm CMOS technologies, which are the target technologies in this work.
As mentioned in chapter 2, for PMOS transistors NBTI takes place when the gate terminal is negatively biased with respect to its source and drain terminals, resulting in the inversion state of the PMOS transistor. Therefore, the stress condition is fulfilled when logic "0" is applied to the gate terminal of the transistor and logic "1" is applied to the source and/or drain terminal of the transistor. After the stress is removed NBTI shows a recovery phenomenon [56, 15]. In other words, immediately after the end of the stress phase, the drift in the threshold voltage has the largest value [51], but partly recovers with a short time constant. By at-speed detection of data transitions, the developed design is insensitive to recovery behavior of NBTI.
During the switching of the transistors, they are exposed to conductive HCI (CHCI). Since disabling the monitors results in absence of any transition, CHCI is avoided.
In Fig. 5.4 the delay element is protected by the Monitor Enable signal. The Monitor Enable signal forces all the NAND gate outputs and data path inputs to the logic "1" and thus removes the stress condition in the data path and the XOR structure. Even though the input PMOS transistors connected to the data and the output of the monitored flip-flop may age, their degradation is negligible compared to the delay of the delay element. When Monitor Enable is deactivated, NAND gates retain the monitor structure in a non-switching state, which reduces the power consumption of the monitor.

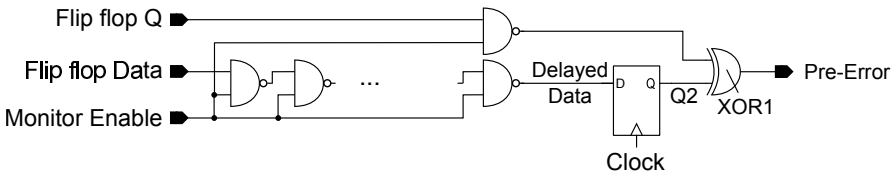


Figure 5.4.: Delay based monitor with Monitor Enable for aging resistance

5.1.2. 2-bit in situ TDC Monitor

One bit monitors distinguish between relaxed and critical transitions by only one threshold. In order to have more precise information regarding the reliability status of the circuit under test the degradation level should be monitored by a higher resolution. Therefore, in this work monitors with more than one detection threshold are designed. The developed mini time to digital converter (Mini TDC) is an in situ monitor with few output states. Fig. 5.5 shows the design of a 2-bit in situ TDC monitor which identifies 4 states for the remaining timing slack. The

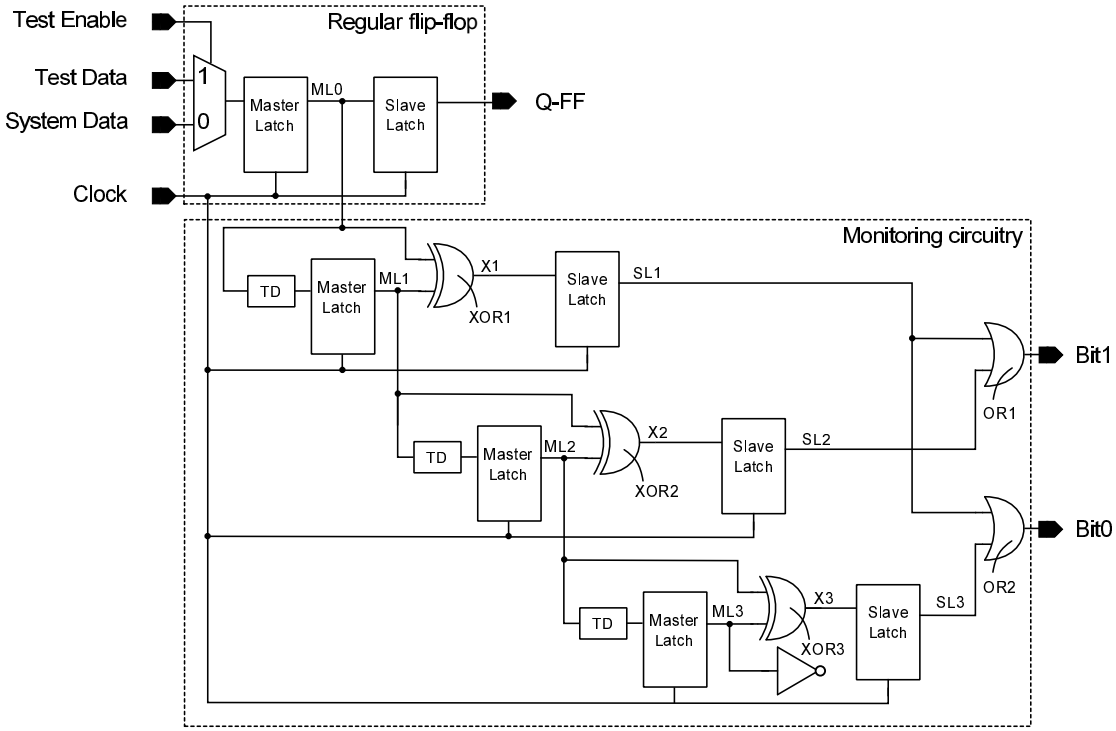


Figure 5.5.: Schematic of the 2 bit in situ TDC monitor, to be placed at the end of the combinatorial path under test. The regular flip-flop is the capturing flip-flop at the end of the path under test.

timing slack is defined as the difference between length of the clock period and the longest propagation delay time between the flip-flops plus the setup-time of the flip-flop.
To minimize the overhead of the monitors compared to the pre-error approach, the output signal of the master latch of a general scan D-flip-flop is utilized. The output of each XOR gate is flagged when equality between the master latch output of the last stage and the current stage is detected. The result of the monitor which is the output of the XOR gate is then latched by the slave latches.
When a relaxed data transition occurs, all of the master latches are able to latch the data by the same clock cycle. This results in $X1\,X2\,X3 = 000$. The resulting output bit word of the monitor in this case would be "00". When the data transition becomes critical, first the master latch 3 fails to latch data, resulting in $X1\,X2\,X3 = 001$ and the monitor output bit word of "01". For data transitions in which master latches 2 and 3 fail to latch the data, $X1\,X2\,X3 = 010$ and the output bit word equals to "10". Similarly for the most critical and yet non-erroneous data transitions in which master latches 1, 2 and 3 fail to latch the data, the result would be $X1\,X2\,X3 = 100$ and the monitor output bit word $(B1\,B0)$ is "11".

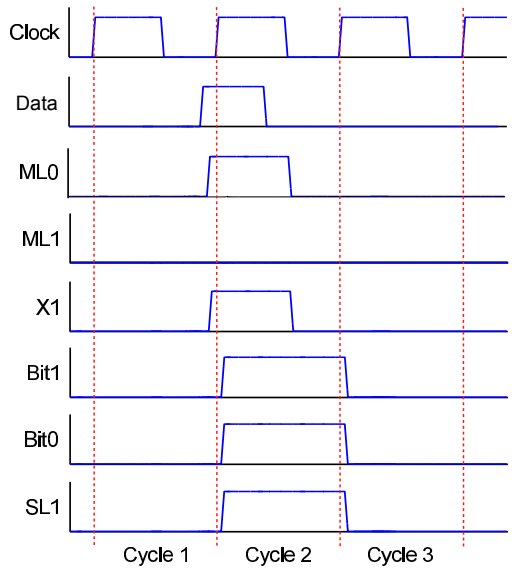


Figure 5.6.: An example of the timing diagram of signals in the 2 bit in situ TDC monitor. In the first cycle a critical transition happens (a data transition close to the triggering edge of the clock). In cycle two master latch 2 (ML2) and ML3 fail to latch the new data and remain at logic "0" resulting in X2 and X3 to remain zero. However, X1 transitions to logic "1". The output of the monitor is the bit word "11" which shows a highly critical and close to error data transition.

Thus, signals $X1$, $X2$ and $X3$ are a one hot representation of the data transition time in reference to the next triggering edge of the clock signal. The one-hot representation is easily converted to a binary representation by the output OR gates. Fig. 5.6 shows the corresponding timing diagram, where a close to error transition occurs, resulting in an output bit word of "11". Fig. 5.7 illustrates the simulation results of the TDC monitor for different data transition times relative to the next clock triggering edge in the corner cases for the 65nm low power technology. The nominal corner is considered as nominal process, supply voltage of $V_{DD} = 1.2\text{V}$ and the temperature of $Temp = 27\,^{\circ}\text{C}$. The slow corner is considered as slow process, $V_{DD} = 1.1\text{V}$ and $Temp = -30\,^{\circ}\text{C}$ (due to the temperature inversion [1]) and finally the fast corner is considered as fast process, $V_{DD} = 1.3\text{V}$ and $Temp = -110\,^{\circ}\text{C}$.

[1]Note that for more mature technologies typically the fast corner is at fast process and low temperature. In advanced technologies, however, the voltage is scaled to a point where the circuit is operated at temperature inversion. Here, the effect of decreasing threshold voltage with temperature exceeds the mobility degradation. Consequently, the circuit exhibits an inverted temperature characteristic, as it speeds up with increased temperature and vice versa.

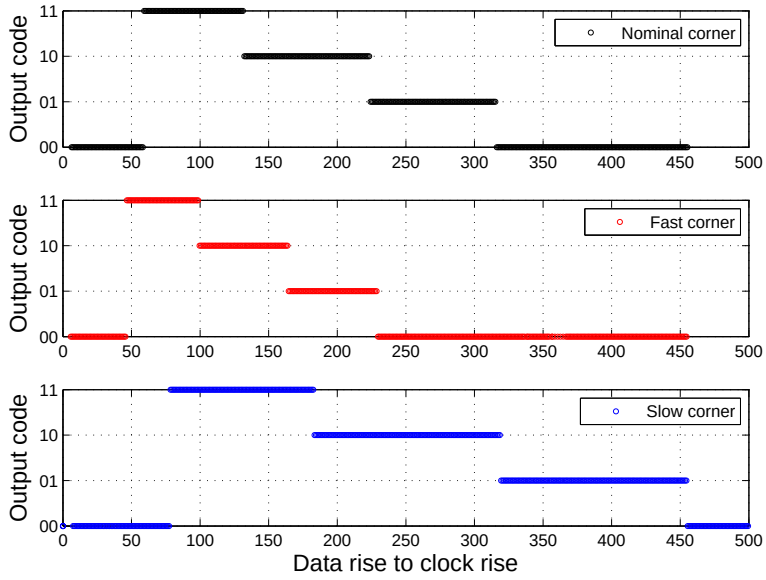


Figure 5.7.: Simulated output of the 2-bit in situ TDC monitor, considering corner cases in the 65nm low power technology.

5.1.3. Precise Slack Monitors

In order to have more precise information regarding the timing properties and thus reliability status of the circuit, it is advantageous to have a time to digital conversion with a high resolution. For this purpose special timing monitors are designed in which the output of the monitor is a pulse with a width of the remaining timing slack of the flip-flop. The output pulse width of the monitor is then converted to a binary code by a sufficiently high resolution time to digital converter.

Low Power Latch Slack Monitor

Fig. 5.8 shows the design of the latch slack generator. When the Monitor Enable is "0", the monitor is disabled. This minimizes the power consumption of the monitor. When the Monitor Enable is activated, occurrence of a data transition during the low phase of the clock results in an output pulse equal to the remaining timing slack of the path under test. In other words the width of the generated pulse is equal to the time interval between the crossing of the data transition and the next rising edge of the clock. The monitor can be turned off by deactivating the Monitor Enable signal. Thus, when the Monitor Enable is logic "0", data transitions do not change the signals within the latch monitor. At this state, the

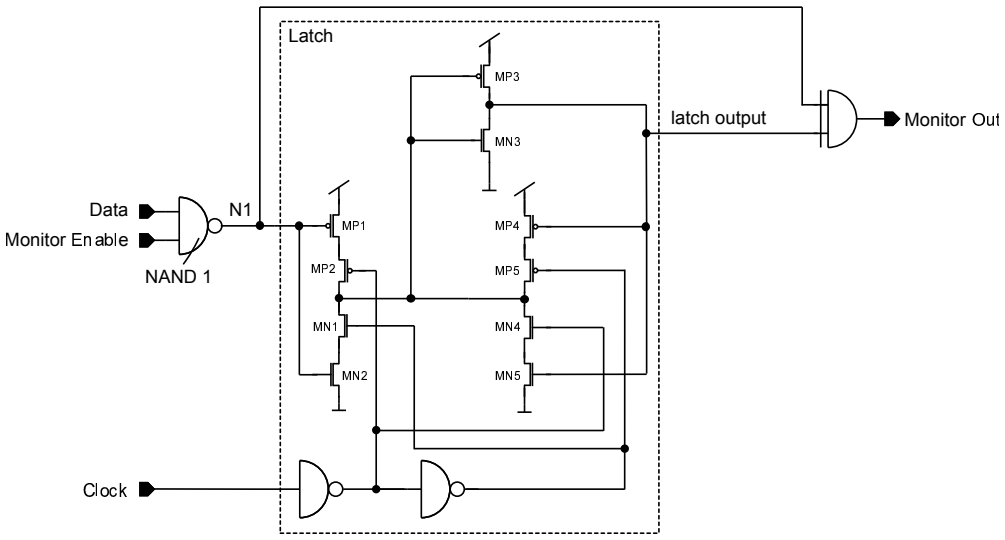


Figure 5.8.: Schematic of the low power latch slack monitor. The internal structure forms a latch which is transparent to its input when the clock signal is logic "1".

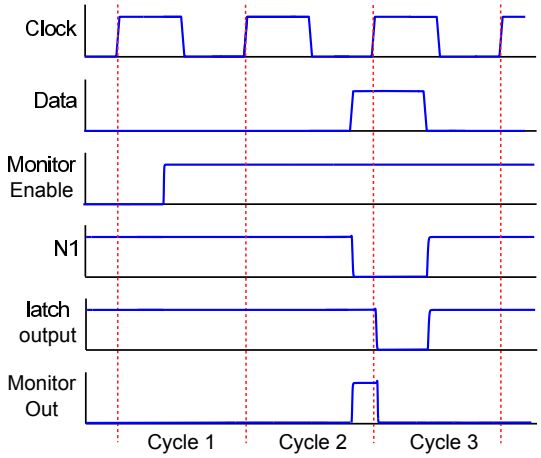


Figure 5.9.: Timing diagram of the low power latch slack monitor

output of the NAND gate is logic "1", resulting in latch output to be also logic "1". Therefore, the XOR gate and first stage of the latch are resistant to NBTI. Compared to the ideal output pulse, the result of the monitor in all corner cases shows low sensitivity to variations (lower than 3ps).
For the supply voltage range of V_{DD} = 1.2V down to VDD=0.9V, the 3sigma interval due to local variations is derived from Monte Carlo simulations and also depicted in Fig. 5.10.

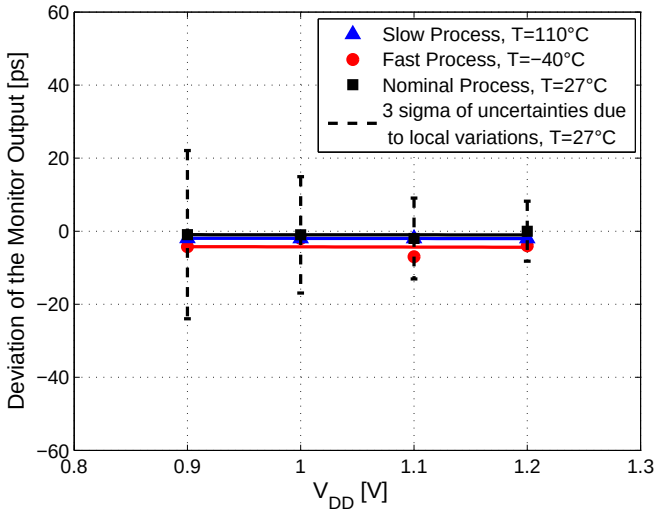


Figure 5.10.: Sensitivity of the low power latch slack monitor to deviations from the ideal pulse width

Aging Resistant Dynamic Slack Monitor

Fig. 5.11 shows the design of the aging resistant monitor based on dynamic logic design [50]. The corresponding timing diagram is shown in Fig. 5.12. In contrast to [93], when a data transition happens during the low phase of the clock, a pulse is generated at the output of the monitor which stays high until the next clock triggering edge. Therefore, the width of the generated pulse is the difference between length of clock period and the propagation delay between flip-flops. In other words, the generated pulse width is equal to the remaining timing slack plus the setup time of the output flip-flop.
The design shown in Fig. 5.11 uses dynamic logic to realize the XOR functionality and detects an inequality between data and delayed data. The monitor uses a Monitor Enable signal to minimize the aging of devices. To remove the stress condition by the Monitor Enable signal, the output of the NAND1 gate is logic "1" when the Monitor Enable signal is logic "0". As a result, the gate terminal of the transistor MP1, the PMOS transistor of the monitor, stays at logic "1", avoiding its aging due to NBTI. When the Monitor Enable is logic "1" the output of the NAND1 gate is the inverted clock, which generates the evaluation phase. Therefore, data transitions generate a pulse at the output of the monitor.
The setup-time of the standard scan flip-flops used in the 65nm technology and the slow corner is approximately 80ps for a supply voltage of $V_{DD} = 1.2$V and is increased to approximately 170ps at $V_{DD} = 0.9$V. If the widths of generated pulses at the output of the monitors are smaller than the flip-flop setup times at the corresponding supply voltages, errors occur which is not acceptable in highly reliable

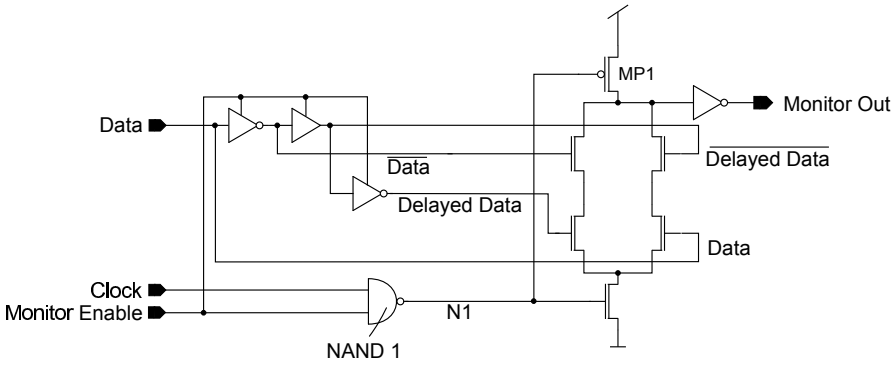


Figure 5.11.: Aging resistant dynamic slack monitor

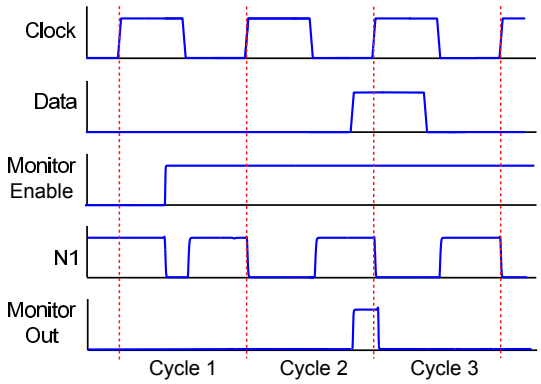


Figure 5.12.: Timing diagram of the aging resistant dynamic slack monitor

circuits. Therefore, for error-free operation of the circuit under test, generated monitor outputs should have bigger lengths than the flip-flop setup times at the corresponding supply voltages. This relaxes the constraint for the minimum time interval to be measured by the time to digital converter, as very small pulses are identified as timing errors and do not have to be measured accurately.
Fig. 5.13 illustrates the deviations of the monitor output in the corner cases from the ideal pulse width in the nominal case. The ideal pulse width is considered as the generated pulse width at the supply voltage of $V_{DD} = 1.2$V, nominal process and a temperature of $Temp = 27\,^{\circ}$C. Monte Carlo simulations are performed to determine the uncertainties of the monitor output due to local variations. For the supply voltage range of $V_{DD} = 1.2$V down to VDD=0.9V, the 3sigma interval due to local variations is derived from Monte Carlo simulations and also depicted in Fig. 5.13. In this design, global variations have a minor impact compared to local ones. As can be seen in Fig. 5.13, the developed timing monitor is robust against supply voltage change. This is desirable for accurate timing measurement of the systems capable of dynamic adaptation of operating parameters especially supply

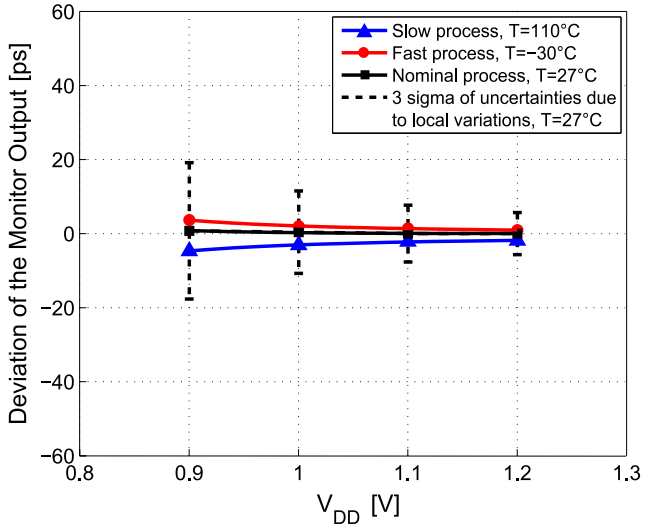


Figure 5.13.: Sensitivity of the dynamic slack monitor to deviations from the ideal pulse width

voltage [90]. Moreover, the monitor structure shows low sensitivity to process and temperature variations.
Compared to the design in standard library elements, the custom designed timing monitor shown in Fig. 5.11 is designed to minimize the degradation of the monitors as well as reducing its power consumption by introducing the Monitor Enable signal. Therefore, the aging resistant timing monitor has a high accuracy considering PVTA variations while it is also optimized in terms of power and area consumption.

Aging Resistant Static Slack Monitor

Figure 5.14 shows the design of the developed aging resistant monitor in static design [6]. Figure 5.15 shows a commonly used master-slave flip-flop equipped with the developed monitor. The corresponding timing diagram is shown in Fig. 5.16. Occurrence of a data transition during the low phase of a clock cycle generates a pulse at the output of the monitor. This pulse stays high until the output of the flip-flop toggles, thus the next clock triggering edge.
For data transitions during the clock high phase, the monitor output rises directly after the clock falling edge. Such a pulse shows a relaxed transition. For data transitions during the clock low phase, the width of the generated pulse has direct correspondence to the time interval between the crossing of the data transition

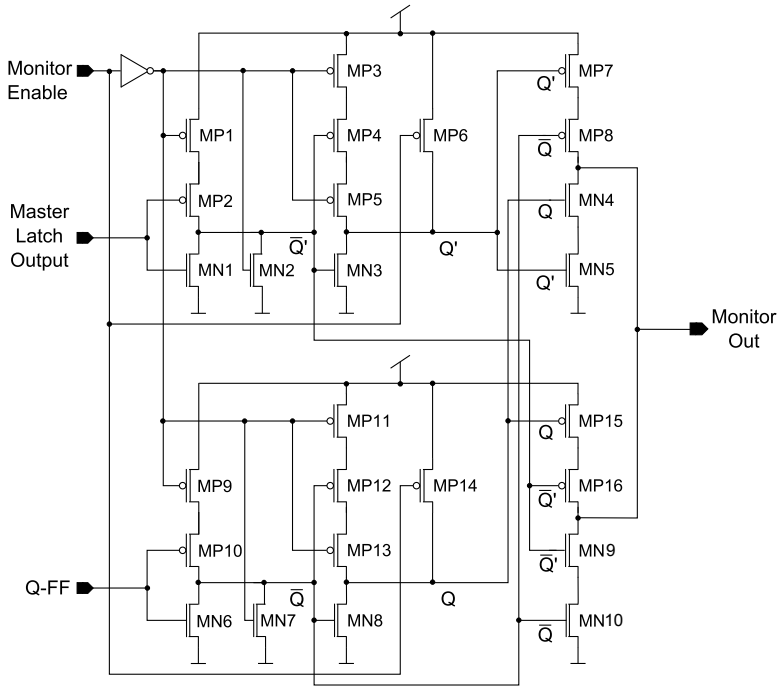


Figure 5.14.: Aging resistant static slack monitor

and the next rising edge of the clock. Using the output of the master latch as the input of the monitor takes the setup time of the flip-flop and its variations into account. In other words, the generated pulse width is equal to the remaining timing slack plus the clock to Q delay of the flip-flop.
The design shown in Fig. 5.14 uses static logic to realize the XOR functionality, detecting an inequality between output of the master latch and Q which is required for transition detection. It also uses a Monitor Enable signal to minimize the aging of monitors. Similar to the dynamic design, the Monitor Enable signal removes the stress condition when disabling the monitors. To minimize the aging of transistors as well as the power consumption of the design shown in Fig. 5.14, an approach similar to cell-based sleep transistor implementation is used [94]. In the cell-based sleep transistor cell approach, a power gating control signal is used to control the sleep transistor. This transistor is used either as a footer device (NMOS) or a header device (PMOS). A weak pull-up/down device controlled by the sleep signal is added to prevent floating outputs in sleep mode.
To protect aging of the PMOS transistors in the first stage in Fig. 5.14, Monitor Enable works as the sleep signal, connected to the header device in the first stage. Therefore, the gate terminals of MP1 and MP9 are connected to $\overline{\text{MonitorEnable}}$ which is logic "1" when monitors are disabled. Nodes $\overline{Q'}$ or $\overline{Q}$ are pulled down to logic "0" by transistors MN2 and MN7, used as pull down devices. Here, drain

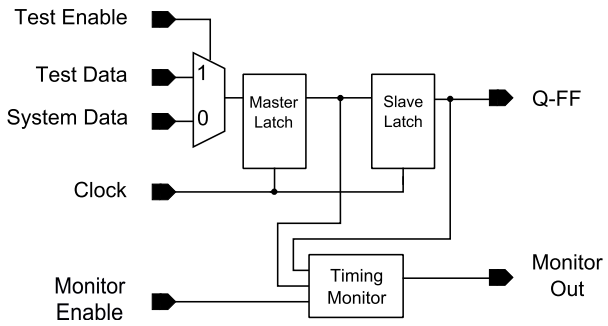


Figure 5.15.: A master-slave flip-flop equipped with the developed aging resistant timing monitor

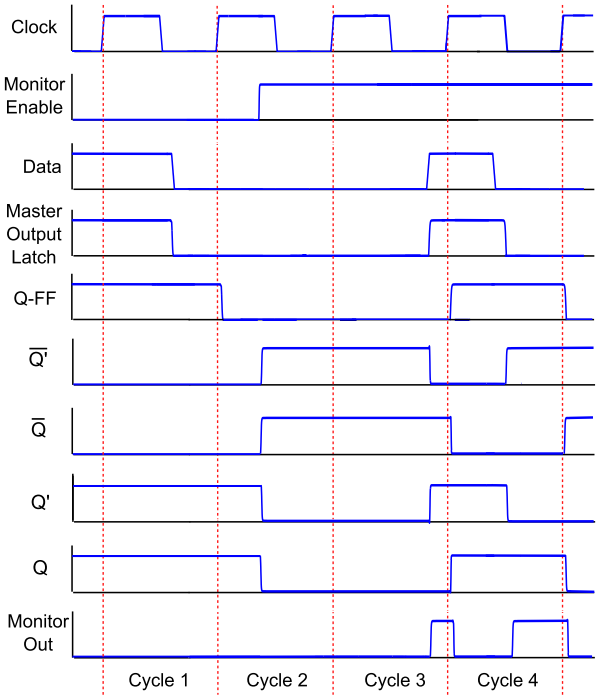


Figure 5.16.: Timing diagram of the aging resistant static slack monitor

terminals of transistors MP2 and MP10 are pulled down to logic “0” and thus these transistors are also protected against NBTI. In the second stage, as signals $\overline{Q'}$ and $\overline{Q}$ are pulled down to logic “0” and are connected to MN3 and MN8, a footer controlled by Monitor Enable is not necessary. However, outputs of the second stage, Q′ and Q, have to be pulled up to logic “1” through transistors MP6 and MP14, protecting PMOS transistors MP7, MP15, MP8 and MP16 in the next stage.
To remove the stress condition for the PMOS transistors of the second stage,

MP4 and MP12, these transistors are isolated by connecting Monitor Enable to the gates of the transistors MP3, MP5, MP11 and MP13. The reason for this is that nodes Q′ and Q are pulled up to logic "1" while the gate of transistors MP4 and MP12 are pulled down to logic "0". Therefore, these transistors would experience stress condition, if they were not isolated from nodes Q′ and Q.
Q′ and Q are then connected to the gates of upper PMOS of the output stage, serving as sleep transistors. The output stage implements the XOR functionality.

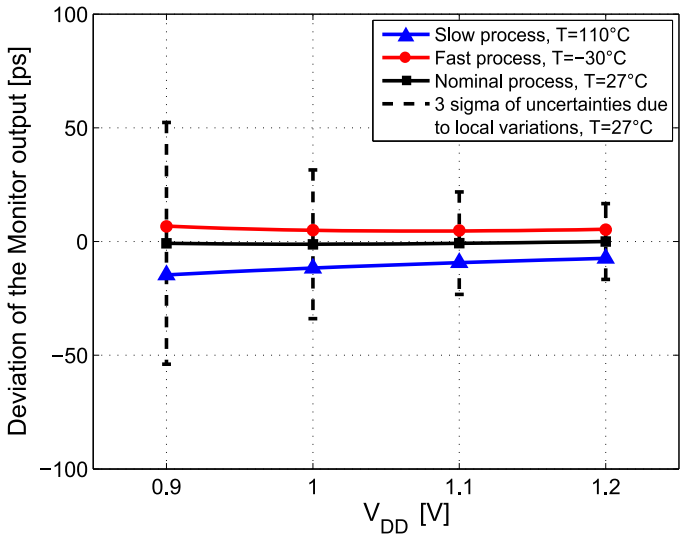


Figure 5.17.: Sensitivity of the aging resistant static slack monitor to deviations from the ideal pulse width

Since the output node "Monitor Out" is pulled down to logic "0" through transistors MN4 and MN5, aging of the PMOS transistors MP7, MP15, MP8 and MP16 is also avoided and a pull down device is not required.
When the clock signal is logic "0" in the circuit shown in Fig. 5.14, the master latch is transparent to data. Therefore, the output of the master latch is equal to data input with a delay. The worst case of this delay determines the setup-time of the flip-flop. Occurrence of a data transition results in an inequality between $\overline{Q'}$ and $\overline{Q}$, which results in a pulse at the output of the monitor.
Setup times of the standard scan flip-flops used in the 65nm technology and the flip-flop equipped with our monitor are evaluated, as shown in Fig. 5.18. Considering the worst case definition for the setup time (obtained in slow process, temperature of $Temp = 110\,°C$) the setup-time of the standard scan flip-flops used in the 65nm technology is approximately increased by 12% and 8% at supply voltages $V_{DD} = 1.2\,V$ and $V_{DD} = 0.9\,V$, respectively. This is small compared to large timing margins added by worst case guard banding approach.
Considering the clock to Q delay, the maximum deviations (slow process, temper-

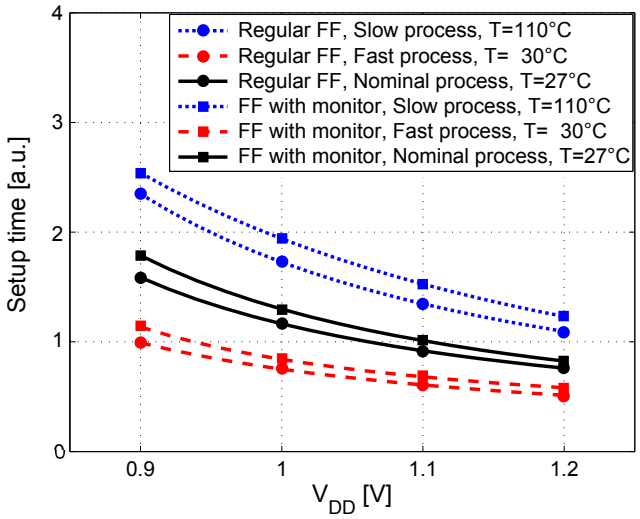


Figure 5.18.: Setup time comparison between standard scan flip-flop and the flip-flop equipped with the aging resistant static slack monitor

ature of $Temp = 110\,°C$) of the regular flip-flop are evaluated as less than 2 ps and 8 ps at supply voltages $V_{DD} = 1.2\,V$ and $V_{DD} = 0.9\,V$, respectively.
Thus, adding monitors to the master slave flip-flop (as in Fig. 5.15) also changes the characteristics of the flip-flops equipped with monitors and thus slightly degrades the performance of the circuit. That means that the setup time and clock to Q delay of the flip-flop slightly increase by adding the monitors. Worst case would be when both launching and capturing flip-flops are equipped with monitors (meaning that the flip-flops at the beginning and the end of a combinatorial path are equipped with monitors). However, at supply voltage of $V_{DD} = 1.2\,V$ and clock frequency of 500 MHz i.e. $T_{clk} = 2\,ns$, worst case performance degradation is less than 1%. At minimum evaluated supply voltage of $V_{DD} = 0.9\,V$ worst case performance degradation is still as low as 1.1%. The deviations of the monitor output in the corner cases from the ideal pulse width in the nominal case is illustrated in Fig. 5.17. The ideal pulse width is considered as the generated pulse width at the supply voltage of $V_{DD} = 1.2\,V$, nominal process and a temperature of $Temp = 27\,°C$. 500 runs of Monte Carlo simulations are performed to determine the uncertainties of the monitor output due to local variations. For the supply voltage range of $V_{DD} = 1.2\,V$ down to $V_{DD} = 0.9\,V$, the 3sigma interval due to local variations is derived from Monte Carlo simulations and also depicted in Fig. 5.17. As can be seen in Fig. 5.17, in this design, global variations have a minor impact compared to local ones. The monitor structure shows rather low sensitivity to process, voltage and temperature variations.

5.2. Abstraction of the Monitor Data

5.2.1. Offline Monitoring

To be able to accurately predict the upcoming failure, it is advantageous to identify the most critical data transition at the end of safety critical paths and provide it to the next processing unit. For this purpose, the output pulses of several monitors are combined into one pulse, which corresponds to the smallest slack in the circuit. This approach also reduces the area and power consumption of the next processing units.
Since critical paths are activated by critical input patterns, each monitor generates a pulse at its output. to determine the most critical delay, it is sufficient to select the narrowest generated monitor output pulse. For this purpose a network with AND functionality is exploited. The network is designed in a way to have the minimum change in the pulse width of the narrowest input. Such a network can be implemented by NAND and NOR gates symmetrical with respect to the inputs, which are shown in Fig. 5.19.
Since the pulse combiner already selects the worst case pulse, a modification of clock duty cycle to tune the detection window of the monitors as proposed in [95] is not required.

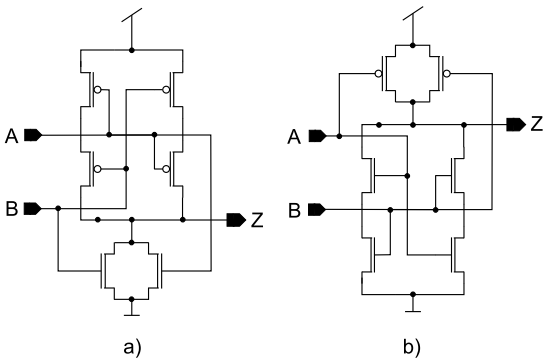


Figure 5.19.: Symmetric gates used in worst case slack selector, 2-input a) NOR gate and b) NAND gate.

5.2.2. Online Monitoring

In an online monitoring system, monitors operate in the functional circuit. As all paths do not transition at their outputs at the same clock cycle, not all monitors generate an output pulse. Thus, each of the monitor output pulses needs to

be measured individually. For an online monitoring system the output of the monitors is combined by a symmetric multiplexer (MUX) tree. Figure 5.20 shows the simulation results for a custom designed MUX tree.

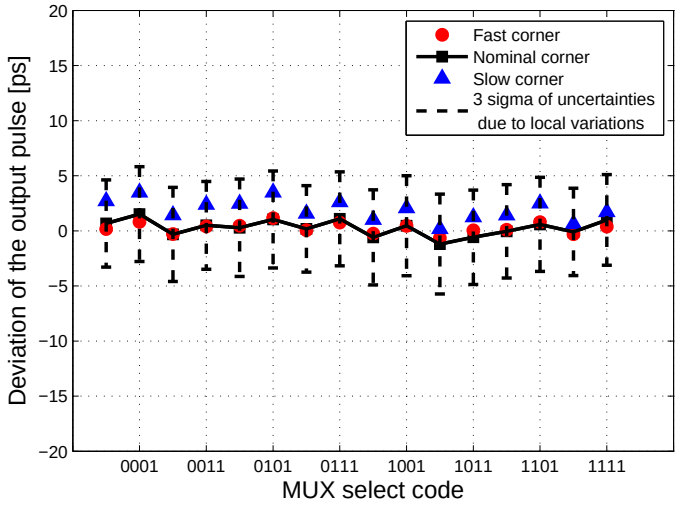


Figure 5.20.: Simulation results of the custom designed MUX tree (16 to 1) including the corner cases and the Monte Carlo simulations considering local variations

5.2.3. Configurability

In order to exploit the defined metadata for accurate prediction of upcoming failures, the pulses generated by the monitors can be tuned based on the reliability requirements. As mentioned before, the pulse width of the output of the monitor represents the remaining timing slack of different sections of the circuit. As the falling edge of the generated pulse at the monitor output is determined by the triggering edge of the clock signal, tuning the rising edge of the monitor output is sufficient. Therefore, a delay element with the ability to delay the rising edge of a signal is used [96]. Fig. 5.21 shows the digitally programmable pulse tuner.
The pulse tuner is optimized for maximum linearity when applying a thermometer code input word. To increase the input bit width, two pulse tuners are connected in series and placed between monitor output and the network which selects the smallest pulse.
A pre-processing unit for 8 monitors MO1 to MO8 with pulse tuning ability is depicted in Fig. 5.22. Tuning values for the monitor output pulse by the pre-processing unit for different states and different voltages is shown in Fig. 5.23a .

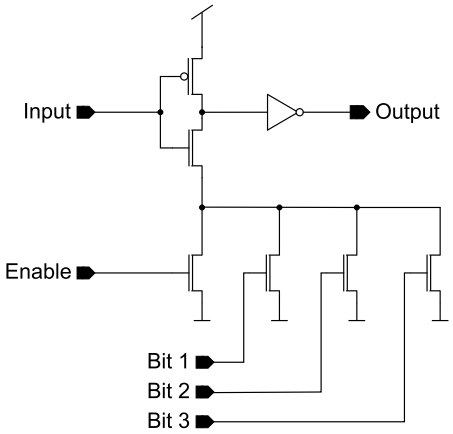


Figure 5.21.: Digitally programmable pulse tuner. The delay element delays the rising edge of a the input pulse.

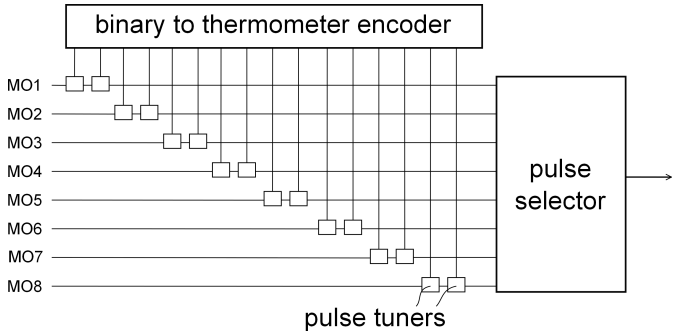
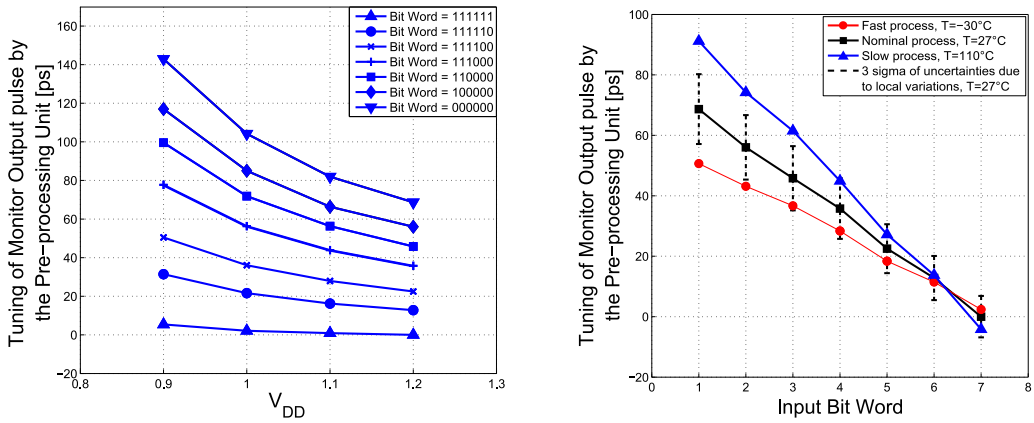


Figure 5.22.: Pre-processing unit with pulse tuning ability for an offline monitoring system selecting the worst case slack of the circuit

The sensitivity of the pre-processing unit with pulse tuning ability for a certain supply voltage of $V_{DD} = 1.2$V in different states is shown in Fig. 5.23b.

5.3. Converting the Timing Slack to the Digital Domain

Time to digital converters (TDCs) measure the time interval between two timing events, the start and the stop signals, and generate a corresponding digital word. The start and stop signals can be assigned as the rising and falling edge of the generated pulse by the timing slack monitors. Therefore, the resulting pulse is applied to a TDC to determine the remaining slack of the entire system as a binary code. The resulting binary code is then transferred to higher layers of abstraction



(a) (b)

Figure 5.23.: a) Tuning values for the monitor output pulse by the pre-processing unit for different states and different voltages and b) Sensitivity of the pre-processing unit for offline monitoring system for a supply voltage of $V_{DD} = 1.2$V. Input Bit Word shows the number of "1" in the thermometer control word.

to diagnose the reliability of the system. For the most reliable test results, the TDC should be implemented close to the circuit under test. Such a TDC should consume low power and area while providing sufficiently high resolution. Moreover, the TDC should be aging resistant.

5.3.1. Delay Line TDC

Figure 5.24 shows a delay line TDC. The start signal propagates through a delay line. When the stop signal is triggered, all outputs of the delay elements are sampled by the sampling flip-flops. Figure 5.25 shows a time measurement based on a reference time window. The measured time interval can be expressed as

$$T_{interval} = T_{out} + \Delta T_{stop} - \Delta T_{start} \tag{5.1}$$

where $\Delta T_{stop}, \Delta T_{start} \in [0, T_{lsb})$. The total error of such an interval measurement would be

$$T_{error} = \triangle T_{stop} - \triangle T_{start} \tag{5.2}$$

Since the start propagation begins by triggering the start signal, ΔT_{start} is zero. Thus, the error is equal to ΔT_{stop}, randomly distributed between 0 and T_{lsb}.
Figure 5.26 shows the results of the corner analysis considering RC parasitics for a 5-bit delay line TDC. The delay elements in the delay line TDC can be buffers

or inverters and the resulting TDC provides a resolution of a gate delay. To have a higher resolution sub-gate delay resolution TDCs can be used. The delay line TDC is not practical for measuring long intervals due to asymmetric layout design, high number of elements and increased area and power consumption. The dynamic range of the delay line TDC can be extended by using a loop structure and a counter. A multiplexer (MUX) can be used for closing the loop.

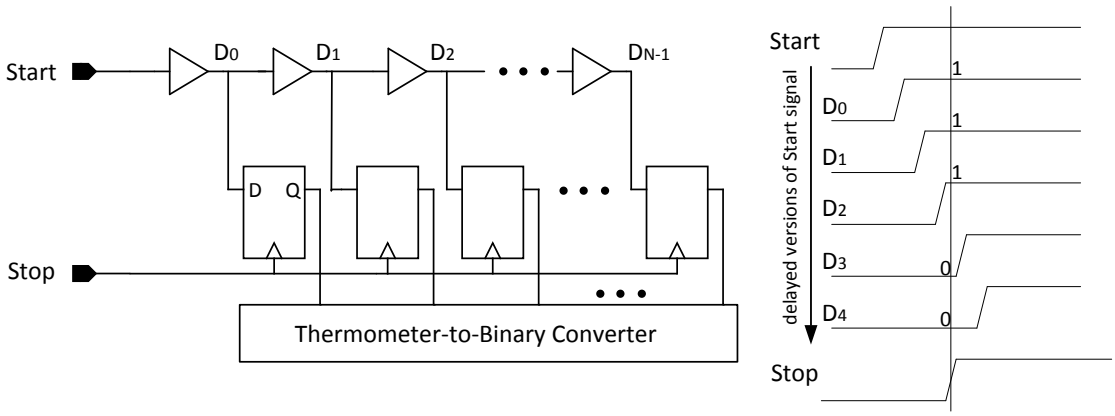


Figure 5.24.: Schematic of a delay line time to digital converter (TDC)

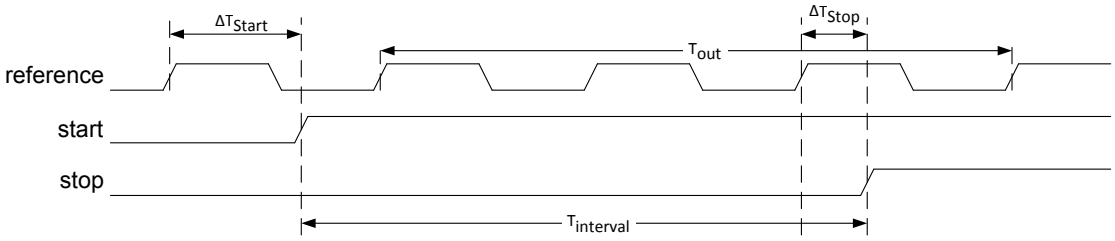
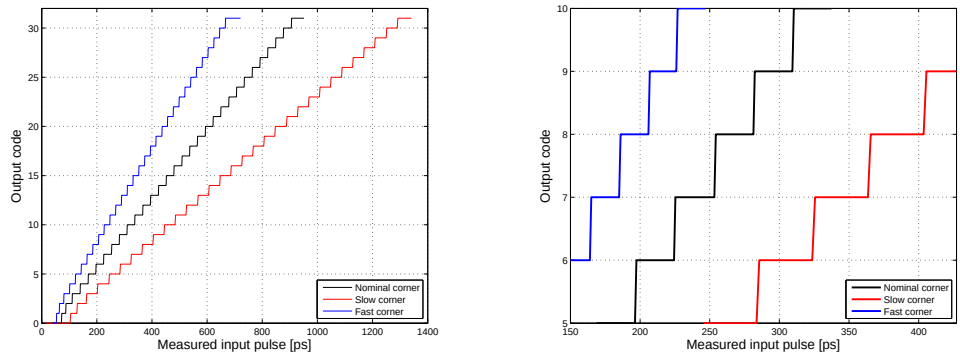


Figure 5.25.: A time measurement by a TDC based on a reference time window, where $T_{error} = \triangle T_{stop} - \triangle T_{start}$

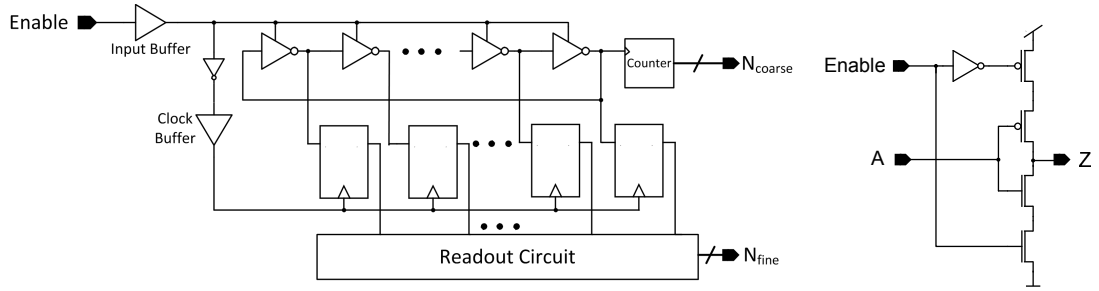
5.3.2. Gated Ring Oscillator TDC

Similar to the delay line TDC, the main core of a gated ring oscillator (GRO) TDC consists of a series of delay elements whose outputs are sampled with the stop event through registers. However, for implementing such a structure current starved inverters as delay elements are used. The start and stop events are the rising and falling edges of the input enable signal, which is the output of the slack monitors. A pulse with the width of the interval which is being measured is applied to the circuit through the enable signal. Such an implementation is depicted in Fig. 5.27. Fig. 5.28 shows the timing waveforms during a measurement.



(a) (b)

Figure 5.26.: Results of the corner analysis considering RC parasitics for a 5-bit delay line TDC, a) full range b) zoomed in characteristics. The T_{LSB} for fast, nominal and slow corners is 20ps, 28ps and 43ps, respectively.



(a) (b)

Figure 5.27.: Structure of a) the basic gated ring oscillator (GRO) TDC and b) inverters with enable/disable capability as delay elements

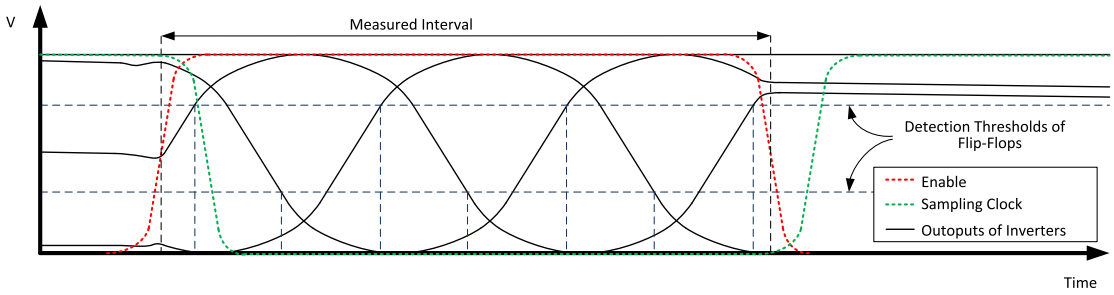


Figure 5.28.: Waveforms of the basic GRO TDC during a measurement

As shown in the Fig. 5.27, an input buffer drives the delay elements. The input buffer is custom designed to minimize the offset error. Note that the signal fed to the clock input of the flip-flops is a delayed and inverted enable signal. This delay eliminates the setup time violations in the sampling flip-flops. Moreover, it avoids double counting the loop while the last inverter in the chain has a transition [97]. Here, the loop counter should be deactivated before the transition of the last inverter in the chain reaches its clock input.
However, this basic structure has several disadvantages as followed. Although the delay of a standard library inverter in 65nm technology can be as low as about 10ps, additional transistors are added in series with the inverter for enabling or disabling the delay element. Thus, it is not possible to obtain a delay below 15ps even with large current starving transistors.
Note that the PMOS current starving transistor in the delay element is turned on or off with one inverter delay after the NMOS transistor. This results in an asymmetry between falling and rising transitions while signal states are held by the disabling signal. Therefore, the gate signal of the PMOS transistors should arrive sooner than that of the NMOS transistors.
Moreover, the conventional master-slave D-flip-flops have a large difference between their detection thresholds for the rising and the falling transitions. In the nominal corner, a rising signal must have a value of more than 750mV to be detected as a logic "1" while for a falling transition this threshold is 350mV. Together with different rise and fall times of the delay elements, this results in a rather big differential non-linearity (DNL) in the characteristics of the TDC.

High resolution multi-path delay elements

In the GRO TDC developed in this work, the interpolation technique is used to halve the resolution of the previously discussed delay element. In this approach, the average of output signals of each two consecutive delay elements is generated and sampled on the arrival of the stop event [98]. The interpolation can be done through resistors or even diodes, which results in high area and power consumption.
Another approach to achieve a higher resolution is to use multi-path ring oscillator [97]. Here, each delay element has more than one input. For each delay element instead of only using the output of the first previous delay element, the outputs of previous delay elements with a distance of an odd number are fed as inputs. For example, the inputs of the 9^{th} delay element in the chain can be the outputs of the 8^{th}, 6^{th}, 4^{th}, 2^{nd}, and so on. Since the output of each previous delay element with a longer distance has its transition sooner than the closer ones, a higher resolution is achieved. However, if the distance is more than half the number of delay elements in the ring oscillator, the farthest signals will begin their transition different to the other ones. For example, if 11 delay elements are present, once the

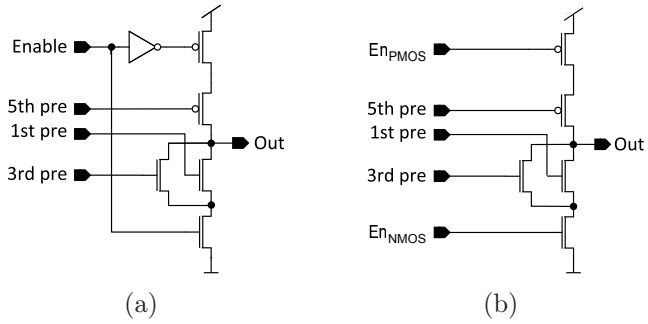


Figure 5.29.: Multipath inverting delay element

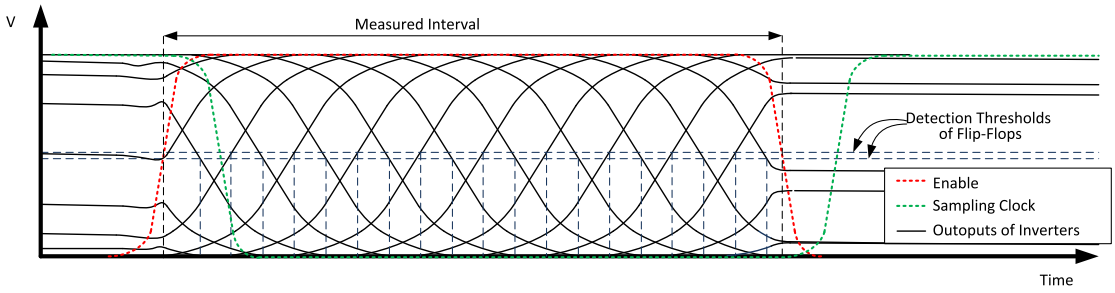


Figure 5.30.: Waveform of the delay element outputs during an interval measurement, with the sense-amplifier flip-flops as the sampling element

9^{th} delay element is having its transition, the 8^{th}, 6^{th} and 4^{th} delay elements are having a transition in the opposite direction. This leads to a transition in the 9^{th} signal. When the 9^{th} signal is having a transition, the 2^{nd} signal also transitions at the same direction as a result of the transition in the 8^{th} signal. Thus, the 2^{nd} signal cannot be used as an input to the 9^{th} delay element. To minimize the area of the TDC, the length of the delay chain is chosen as 15 elements while each delay element has 3 inputs from previous ones. Fig. 5.29 shows the design of the delay element. The farthest signal which has its transition sooner is fed into the gate of a PMOS transistor as it is slower than the NMOS ones.
The waveforms of the outputs of the delay elements during a measurement interval are shown in Fig. 5.30. The difference in the detection thresholds of the flip-flops is reduced by careful design, as explained in the following.

Sampling flip-flops

The conventional master-slave D-flip-flops have large difference between their detection thresholds for rising and falling transitions, as shown in Fig. 5.28. Therefore, sense amplifier flip-flops (Fig. 5.31) are designed with the minimum difference

in the detection thresholds [99]. In such flip-flops, if the data-to-clock delay is large enough, the difference between thresholds in a 65nm technology would be below 10mV in the nominal corner. However, in order to have no setup-time violations, there is a delay of about 100ps to 150ps between holding the analog states of the delay elements and rising edge of the sampling clock signal, t_{D2CP}. The relatively

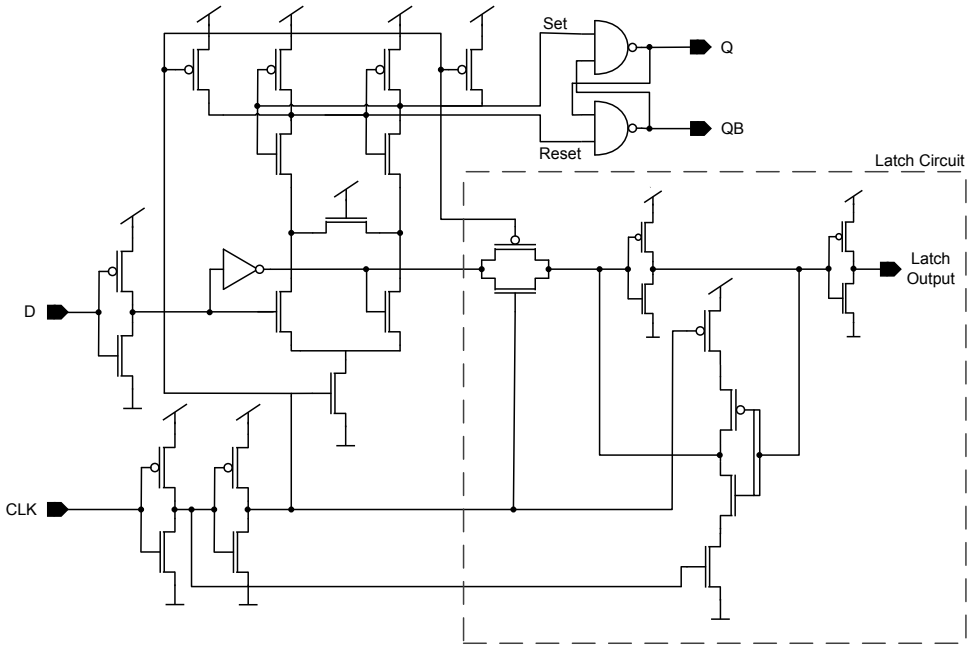


Figure 5.31.: Sense-amplifier flip-flop as the sampling element in the GRO TDC

low delay between the data and the sampling clock edge might lead to an increase in the difference of detection thresholds up to 30mV. This is negligible compared to that of conventional D flip-flops (400mV in the nominal corner, Fig. 5.28). In order to count the number of completed loops, a latch is required for the output of the last delay element in the chain. An additional latch cannot be inserted in parallel to the flip-flop since this will increase the capacitive load of the output of the delay element leading to an asymmetry in the delays of the elements of the ring oscillator. A solution would be to insert additional latches for all outputs of the delay elements. However, this increases the circuit area significantly and decreases the resolution by increasing each delay element load. Another solution would be to add the latch inside the structure of the flip-flops only for the required nodes. By properly sizing the input inverter, both thresholds are shifted up or down together.
When the falling edge of the clock signal is triggered, the equalizer transistor prepares the flip-flop for the next clock rising edge. If the low phase of the clock signal is not long enough, the flip-flop does not work properly in the next sampling

instance. Thus, it must be guaranteed that the equalization time is provided for the flip-flops in any case. Since the outputs of the delay elements can be sampled after their analog states are held constant by the stop event, the rising edge of the clock is delayed up to a certain level which provides enough time for the equalization. Such an operation is performed inside a clock buffer.

Readout circuitry

In addition to the TDC core, a processing circuit is required to convert the result of the TDC core to a binary number representing the measured interval, as shown in Fig. 5.32. When the delay elements used in the core of the TDC are inverters,

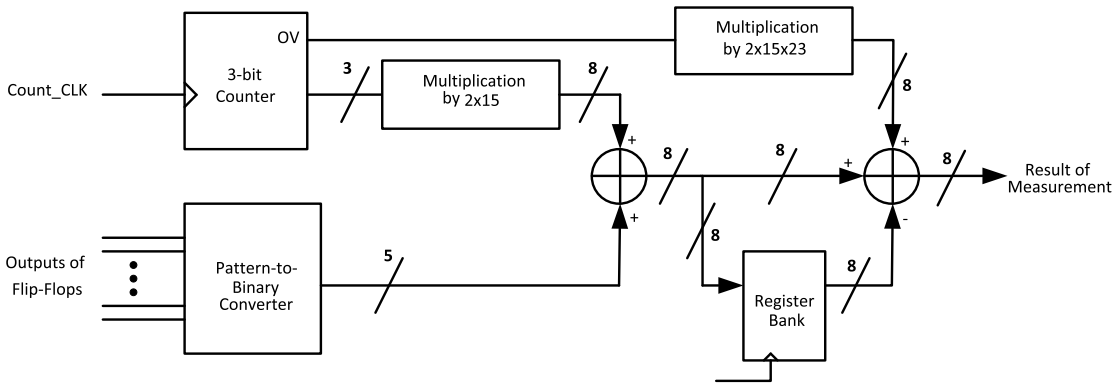


Figure 5.32.: Structure of the readout circuitry for the GRO TDC

the propagating signal is inverted after every stage and the output code sampled by the flip-flops is an alternating sequence of zeros and ones. Thus, the output code follows the pattern of a pseudo-thermometer code. A pseudo-thermometer to binary converter (PTBC) converts the resulting output of the flip-flops ($FF1$, $FF2$, ..., $FF15$) to a 4-bit binary code ($Q3$, $Q2$, $Q1$, $Q0$), denoted as N_{fine} in Fig. 5.34. A change of phase of the alternating sequence indicates the length of the measured interval. In order to detect the last transition in the chain, every bit should be compared to the next. Thus, in the first step the generated pattern is converted to a one-hot code. A one-hot code corresponds to a combination of digital bits, of which only a single bit is logic high and all the others are logic low. Moreover, a first-order bubble correction is performed to obtain more reliable results. Thereby, a bit within a pseudo-thermometer pattern is compared to both its preceding and its following bit. Afterward, the one-hot code is converted to a binary code by using an encoder, as shown in Table 5.1 and Fig. 5.33.
In case of a input pulse width longer than that of the chain of delay elements, a ring oscillator structure is active, by connecting the output of the last element back

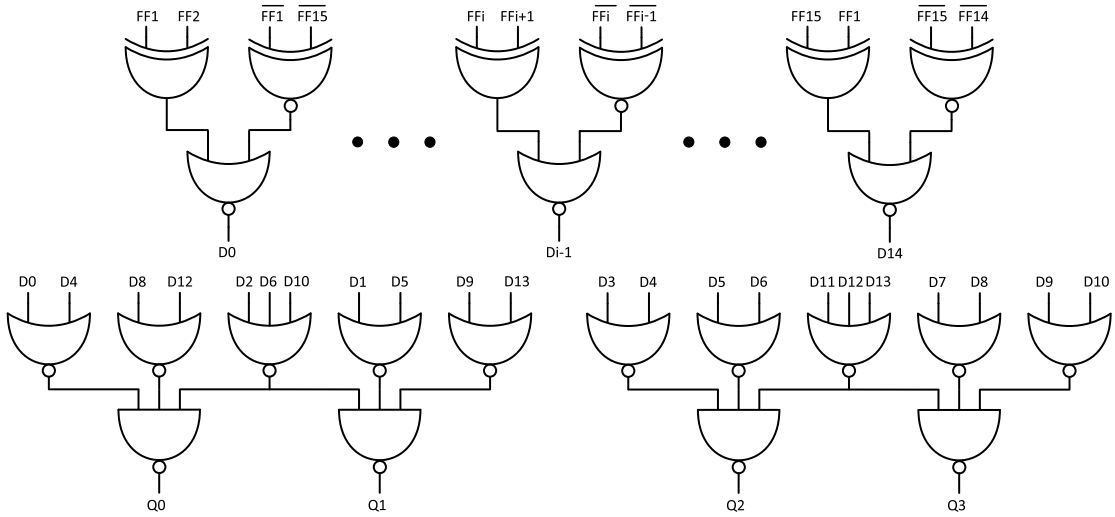


Figure 5.33.: Pseudo-thermometer to binary converter (PTBC) converts the outputs of the sampling flip-flops to a binary code

to the input. The number of full oscillations is counted by a custom designed 3-bit counter. This counter is enabled during the interval being measured and counts the rising transitions of the last delay element in the chain.
After the falling edge of the enable signal and the sampling through the flip-flops, an extra circuit detects the last signal in the chain with a transition. This along with the sampled value of the last signal in the chain determines the phase state of the ring oscillator at the disable instance. If there are N_d delay elements in the ring oscillator, there would be $2\,N_d$ different phase states. Afterward, $2\,N_d$ times the counter value is added to the delay value from the chain. Then the result is subtracted from the previous interval measurement. As the number of delay elements $N_d = 15$, the result of the counter is multiplied by $2\,N_d = 30$ and added to the binary output corresponding to the phase state provided by PTBC. To multiply the counter value by 30, it is easier to subtract twice from 32 times the counter value. The circuit for which the delay monitor is used, has a frequency of 500MHz, meaning that the dynamic range of the TDC needs to be 2ns. If the resolution of the TDC is 10ps, each full oscillation takes $2 \cdot 15 \cdot 10 = 300$ps. Thus, a 3-bit counter can measure a TDC dynamic range of 2ns. The output of the counter is , denoted as N_{coarse} in Fig. 5.34. The deglitch circuit in Fig. 5.34 ensures that the counter clock becomes "1" only when both $O13$ and $O15$ are "1" and becomes "0" when both are "0". Thus, the capacitive load of $O15$ is reduced and the possibility of double counting by the counter is removed.
The PTBC uses the output of the 15^{th} flip-flop ($FF15$) to determine whether it is required to add 15 or not. If $FF15$ is logic one, the PTBC output is the phase state. However, if $FF15$ is logic zero, 15 units must be added to the PTBC result

	Flip-Flops output pattern (Pseudo-thermometer) FF1,FF2, . . . , FF15	One-hot code D14, D13, . . .,D0	Binary output Q3, Q2, Q1, Q0
0	101010101010101 or 010101010101010	100000000000000	0000
1	001010101010101 or 110101010101010	000000000000001	0001
2	011010101010101 or 100101010101010	000000000000010	0010
3	010010101010101 or 101101010101010	000000000000100	0011
4	010110101010101 or 101001010101010	000000000001000	0100
5	010100101010101 or 101011010101010	000000000010000	0101
6	010101101010101 or 101010010101010	000000000100000	0110
7	010101001010101 or 101010110101010	000000001000000	0111
8	010101011010101 or 101010100101010	000000010000000	1000
9	010101010010101 or 101010101101010	000000100000000	1001
10	010101010110101 or 101010101001010	000001000000000	1010
11	010101010100101 or 101010101011010	000010000000000	1011
12	010101010101101 or 101010101010010	000100000000000	1100
13	010101010101001 or 101010101010110	001000000000000	1101
14	010101010101011 or 101010101010100	010000000000000	1110

Table 5.1.: One-hot and binary codes corresponding to each pattern at the outputs of the flip-flops in the GRO TDC

to achieve the phase state. For such an operation, the output of the PTBC is added to 4 repeated bits of $\overline{FF15}$.

Final structure and results

The final structure of the GRO TDC is illustrated in Fig. 5.34. Figure 5.35 shows the characteristic diagram of the GRO TDC. Figure 5.36a shows The differential non-linearity (DNL) in terms of LSB zoomed in for the maximum values along all output codes considering the corner cases and Fig. 5.36b shows the result of Monte Carlo simulations considering local variations for arbitrary measured time intervals. Figure 5.37 shows the power consumption in the nominal case.

5.3.3. Aging Resistant NAND Gate TDC

To decrease the area and power consumption of the TDC while adding the aging resistance feature, an aging resistant ring oscillator TDC constructed by NAND gates is developed, as shown in Fig. 5.38. The main core of the developed ring oscillator TDC consists of a series of NAND gates as delay elements. The outputs of the delay elements are sampled with the stop event through conventional master-slave D-flip-flops as sampling components. The second inputs of all NAND gates

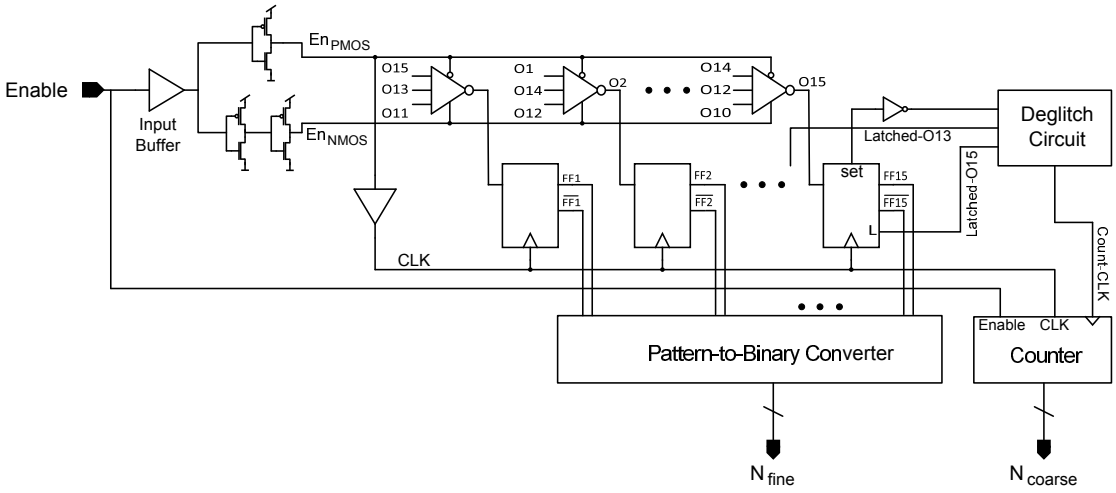
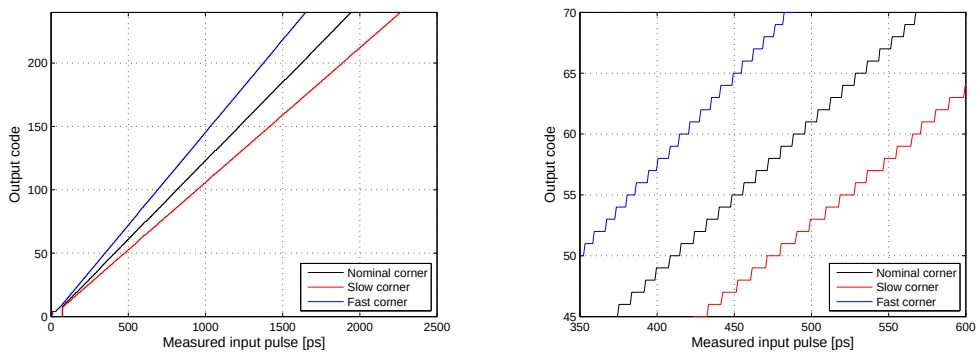


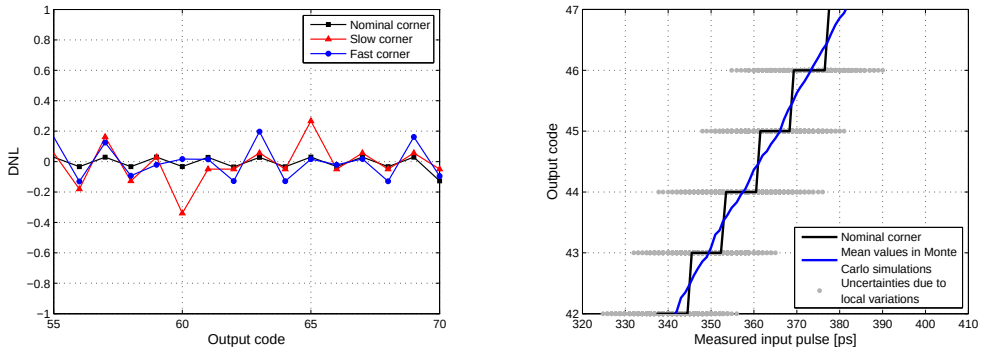
Figure 5.34.: Developed structure of the GRO TDC



(a) (b)

Figure 5.35.: GRO TDC characteristic diagram for different corner cases: a) full range and b) zoomed in. The T_{LSB} for fast, nominal and slow corners is 6.7ps, 8.0ps and 9.5ps, respectively.

are connected to the Monitor Enable signal to minimize the aging of the TDC circuit. When the monitoring is disabled, i.e. Monitor Enable is logic "0", all NAND gates have a logic "1" at the output, as shown in Fig. 5.39a. This prevents aging of the PMOS transistors inside the Ring Oscillator due to NBTI. However, PMOS transistors of the NAND gates connected to Monitor Enable age during disabled monitoring time. Nevertheless, these PMOS transistors are not inside the loop of the ring oscillator. Therefore, aging of these transistors does not affect the operating speed of NAND gates inside the ring oscillator and thus does not affect the resolution of the TDC.
The PMOS of the NAND gate connected to the Enable signal also ages. However,



(a) (b)

Figure 5.36.: a) GRO TDC differential non-linearity (DNL) in terms of LSB zoomed in for the maximum values along all output codes considering the corner cases b) Result of the Monte Carlo simulations considering local variations for arbitrary measured time intervals

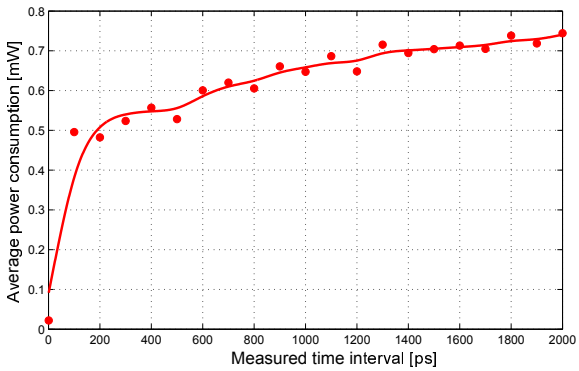


Figure 5.37.: Average power consumption of the GRO TDC for different measurement intervals

since the start event is triggered by the rising edge of the Enable signal, aging of this PMOS transistor also does not affect the operation of the TDC.
When Monitor-Enable is activated (logic "1"), the interval which has to be measured is given to the TDC as the Enable signal, i.e. start and stop events are considered as the rising and falling edges of the enable signal. As mentioned before, the Enable signal is the output of an in situ monitor within and corresponds to the worst case slack of the circuit under test. Before the rising edge of the Enable signal, the output of the first NAND gate is logic "1" and all other NAND gates have either logic "1" or logic "0", as shown in Fig. 5.39b. Thus, all NAND gates have constant outputs and the oscillator has a stable state.

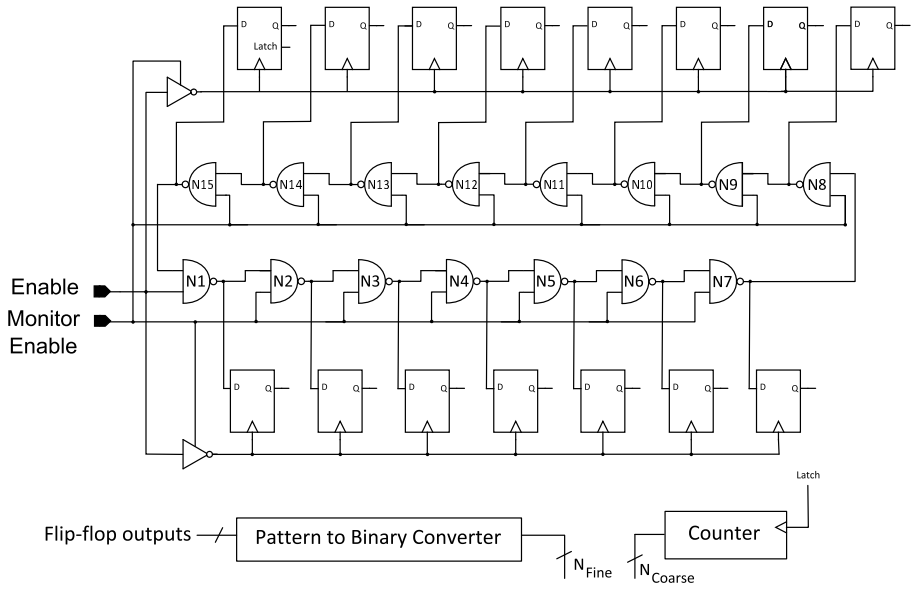


Figure 5.38.: Aging resistant ring oscillator TDC with NAND gates

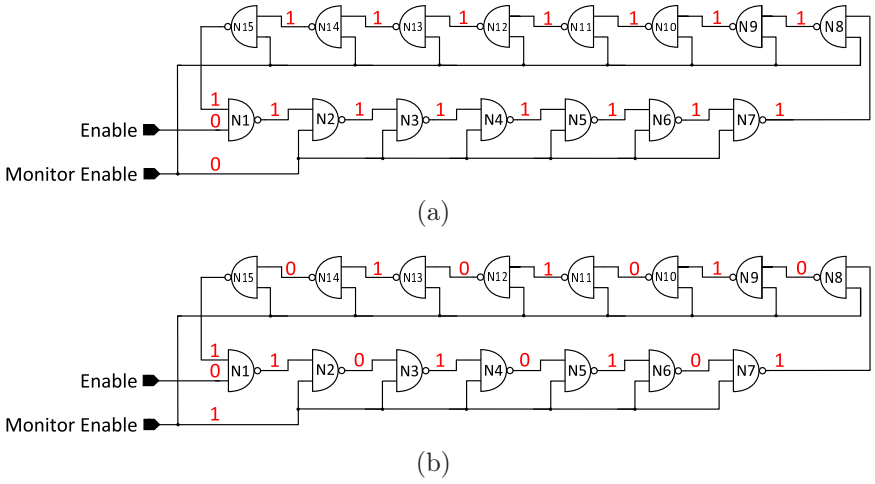
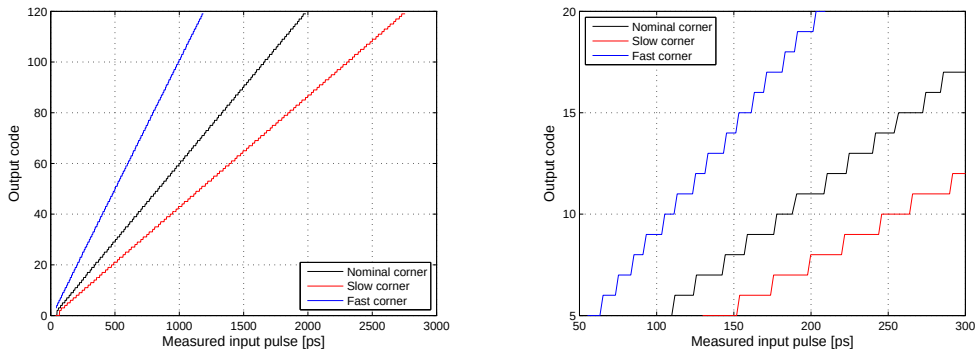


Figure 5.39.: Aging resistant NAND loop, a) deactivated Monitor-Enable and b) activated Monitor-Enable and NAND gates in stable state

After the rising edge of the Enable signal, NAND gate N1 works as an inverter. Therefore, the transition is passed through the NAND gates and the oscillator begins to oscillate. By the falling edge of the Enable signal, the oscillator attains a stable state and the number of gates that were passed by the start event is determined. This number has a direct correspondence to the length of the measured interval. The resolution of such a TDC is equal to the delay of each NAND gate. However, since the detection thresholds of the flip-flops for the rising and falling

edges are different, the resolution of the TDC is limited. Considering only rising transitions and a simple NAND ring oscillator, the resolution would be twice the delay of NAND gates.
To measure intervals longer than the oscillation period, a counter counting the number of full loop oscillations during operation is used. Since outputs of $N_d = 15$ NAND gates are available, $2\,N_d = 30$ different phase states are obtained, similar to the GRO TDC (section 5.3.2). Each full oscillation takes 30 times the delay of one NAND gate. The delay of a NAND gate in this structure and in 65nm technology is evaluated as 16.4ps, resulting in full oscillation of 492ps. Such a resolution is sufficient for slack measurement of a system with clock frequency of 500 MHz. A transition within a clock cycle is measured using a 2-bit counter. Figure 5.40 shows the characteristic diagram of the aging resistant NAND TDC. Figure 5.41a shows The differential non-linearity (DNL) in terms of LSB zoomed in for the maximum values along all output codes considering the corner cases. Figure 5.41b shows the result of the Monte Carlo simulations considering local variations for arbitrary measured time interval. Figure 5.42 shows the power consumption of the aging resistant NAND TDC in the nominal corner.

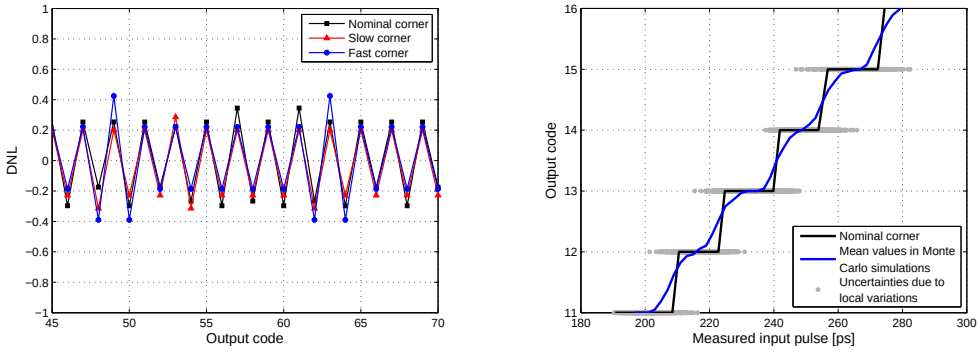


(a) (b)

Figure 5.40.: Aging resistant NAND TDC characteristic diagram for different corner cases: a) full range and b) zoomed in. The T_{LSB} for fast, nominal and slow corners is 9.8ps, 16.4ps and 23.3ps, respectively.

5.4. Overhead of the Monitoring System

To evaluate the overheads of the monitoring system, timing monitors are integrated into a circuit with an array of eight 16-bit multipliers synthesized with an industrial design flow in a 65nm CMOS technology. Monitors are placed at the



(a) (b)

Figure 5.41.: a) Aging resistant NAND TDC differential non-linearity (DNL) in terms of LSB zoomed in for the maximum values along all output codes considering the corner cases b) Monte Carlo simulations considering local variations for arbitrary measured time interval

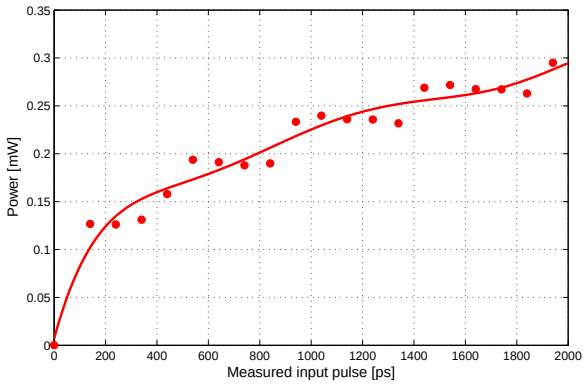


Figure 5.42.: Average power consumption of the aging resistant NAND TDC for different measurement intervals

end of the eight most critical paths determined by SPICE simulations, 3.1% of all 8×32 outputs.

2-bit in situ TDC Monitors

In situ TDC monitors provide the remaining timing slack as a 2bit digital code. The power overhead of the monitors for the above mentioned arithmetic circuit would be as low as 3.0%.

Dynamic Slack Monitor in Combination with the GRO TDC

The outputs of the precise slack monitors are fed to the pre-processing unit and then the resulting pulse to the time to digital converter. The clock frequency is chosen as 500 MHz ($T_{clk} = 2$ ns). When analyzing the power overhead of the monitoring approach, the overhead of the entire monitoring circuitry has to be considered. Additional power consumption arises from of the extra circuitry for the timing monitors, pre-processing unit and the TDC. During normal operation when the monitoring circuits are disabled the power overhead of the timing monitors and the pre-processing unit are negligible (together less than 0.4%). The power overhead of the TDC is also small and evaluated as 0.3%. Therefore, the power overhead of the entire monitoring system when disabled is less than 0.7%. During monitoring by applying the predetermined worst case patterns, the power overhead of the timing monitors is evaluated as (0.82%), the power overhead of the pre-processing unit is 0.78%. The power overhead of the TDC in this case is 11.1%. Finally, the total power overhead when applying the worst case patterns is 12.7%. For large circuits equipped with the monitors only at few paths, the power overhead of the monitoring system is reduced to a great extent, as the main contributor to the power overhead is the TDC.

Static Slack Monitor in Combination with Aging Resistant NAND TDC

Here the situation is somewhat different. When disabling the monitoring circuits by the Monitor Enable signal, the power overhead of the entire monitoring system is negligible and drops to less than 1%. When monitoring by applying the worst case patterns, the power overhead of the timing monitors is evaluated as less than 1%. The power overhead corresponding to the pre-processing unit is also less than 1%. The total power overhead considering in situ monitors, the pre-processing unit and the TDC is only 6.1%. Here, the main contributor to the power overhead is the TDC, with a power overhead of 4.5%.

5.5. Summary

In this chapter the required circuitry for precise monitoring of timing properties was discussed. Design and properties of different developed in situ monitors were discussed and evaluated. The timing information extracted by the monitors are converted to the digital domain by either a decentralized approach (output of the monitors is a binary value) or by a centralized approach utilizing time to digital converters.

6. Evaluation of the Monitoring System in Medical Applications

Increased supply voltage in over-constrained designs with large guard-bands accelerates the aging of the circuitry during the lifetime and results in waste of power, area and performance for the digital circuit. Thus, for low power and strictly reliable applications, too extreme guard banding of operating parameters should be avoided. In such applications advanced methods such as adaptive voltage scaling (AVS) are of interest.
In medical implants such as neural measurement systems (NMS) [100, 101] reliability in combination with power efficiency is a crucial design goal. Such systems need to operate error-free during rather long lifetimes up to decades (e.g. 30 years). Thus, high reliability requirements need to be met and performance degradation through aging of the circuitry should be avoided. Moreover, delivering a high power to the implanted chip within the body is very difficult. High power consumption may result in a higher than body temperature for the implant. This leads to health risks and hazardous pain for the patients. Therefore, for such applications both reliability and power consumption become important design criteria.
In the monitoring approach [7] presented in this chapter, in situ timing monitors are inserted and fabricated within the digital front end of an NMS. The monitors can distinguish between critical and relaxed operation, as discussed in section 5.1.1 [90, 91]. Extracted timing information is used for on-line adaptation of the supply voltage in order to reduce the power consumption as well as the device aging in the implants. Thus, the supply voltage of the digital circuitry is decreased to the minimum possible value. This is possible by enabling the observability of the circuit's performance and ensuring reliable operation of the circuit.
In this chapter, implementation of the reliability monitoring approach in combination with adaptive voltage scaling for a neural measurement application is discussed. The digital front end of an NMS is equipped with the in situ timing monitoring system and the applicability of the monitoring approach for the NMS is evaluated by experimental data. All circuits presented in this chapter are designed and fabricated in a 350 nm CMOS technology with a nominal supply

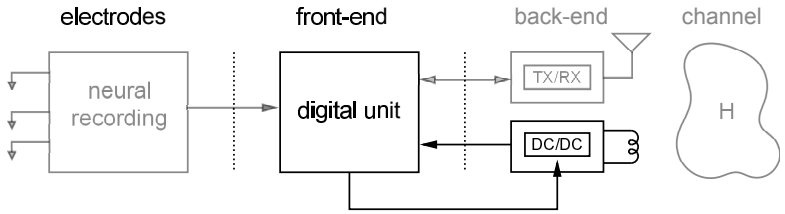


Figure 6.1.: The neural measurement system [100, 102] is equipped with the monitoring system. The arrow from the digital unit to the DC/DC converter represents the extracted and pre-processed monitor data.

voltage of $V_{DD} = 3.3\,\mathrm{V}$.

6.1. Circuit under Test

The monitoring system is implemented within the digital front-end of an NMS [100, 102], which is implanted into the human brain. The NMS measures spikes and local field potentials in the brain. It includes measuring and pre-processing of the neurological data which are transmitted over a wireless RF datalink [102]. The neural signals in the brain are recorded by passive electrodes or needles. The signals are pre-amplified and passed to the digital front-end. In the NMS, the digital front-end enables processing of neural measurement data, generates an adapted transmission package and includes an interface to the transceiver. The operating power of the NMS is delivered by an inductive energy supply, as shown in Fig. 6.1. To save area and power while counteracting reliability threats, the digital front-end is equipped with the in situ monitors similar to section 5.1.1. The monitors observe the status of circuit-level timing properties and detect the impact of PVT variations plus degradation mechanisms over lifetime. The information extracted by the monitors is pre-processed and passed to the DC/DC converter. The integrated DC/DC converter in Fig. 6.1 is controlled by the results of the monitoring system. The closed loop configuration enables reliable and power efficient operation. The entire monitoring system in the NMS was synthesized within the already existing design flow, which simplifies the integration of the monitoring system within an existing design.

6.2. Monitoring system

As mentioned before, to address both reliability and power efficiency in the neural measurement system, the supply voltage of the digital front-end is reduced. A lower supply voltage reduces the aging and power consumption and increases the

lifetime of the implants within the body.
Monitoring the timing of digital circuits during lifetime enables predicting performance failures. This enables to take countermeasures such as maintenance before failure or alternatively adaptation of operating parameters. Here, for a medical implant to avoid the necessity of maintenance actions, adaptive methods are desirable as maintenance would require substituting the chip in the body.
According to the current status of the circuit, the operating parameter (i.e. supply voltage) is adapted to an optimal value for specific reliability requirements, e.g. a certain error rate. By dynamic adjustment of operating parameters such as supply voltage, over-constrained guard-bands are reduced. Thus, area and power is saved.
To decrease the complexity and avoid occurrence of errors, in contrast to error detection methods [103, 93], the pre-error monitoring approach with the one-bit monitors (section 5.1.1) is used. The pre-error approach utilizes in situ delay monitors capable of detecting critical transitions by only one threshold. As mentioned in section 5.1.1, critical transitions detected by the monitors indicate a reduced timing slack and thus performance degradation due to voltage scaling and/or aging.
As discussed in section 4.2, monitoring the circuit's timing properties is possible during normal operation (functional circuit) or during test sequences. Monitoring during normal operation requires no interference with the operation of the circuit and thus avoids the idle times. In the NMS, the monitoring approach needs to react on the fly to the performance degradation due to aging or a decreased supply voltage. Thus, the on line monitoring is utilized which provides timing information according to the current operating conditions and enables on the fly adaptation of operating parameters.
The timing information regarding the current status of the circuit is provided by in situ delay monitors at critical positions, e.g. end of most critical paths (section 4.2.1). Based on this information, a closed control loop adapts the supply voltage. The in situ delay monitors observe the timing of the circuit, which is affected by PVTA variations. Fig. 6.2 illustrates the timing behavior of the used CMOS technology considering an inverter chain (ring oscillator). In Fig. 6.2 the delay of an inverter chain is simulated by corner analysis. The measured values correspond to measurement results of an integrated ring oscillator within the digital front-end.
The timing information extracted by the monitors is gathered and encoded by an intermediate module. In the next step, this information is transferred as a digital code to the voltage regulator for online adaptation of supply voltage according to the extracted timing information and thus reliability status. Therefore, by the monitoring approach, the supply voltage is adapted over time according to the reliability properties of each individual chip and the power consumption is reduced.
The monitor output signals, i.e. the pre-errors, are generated when the timing

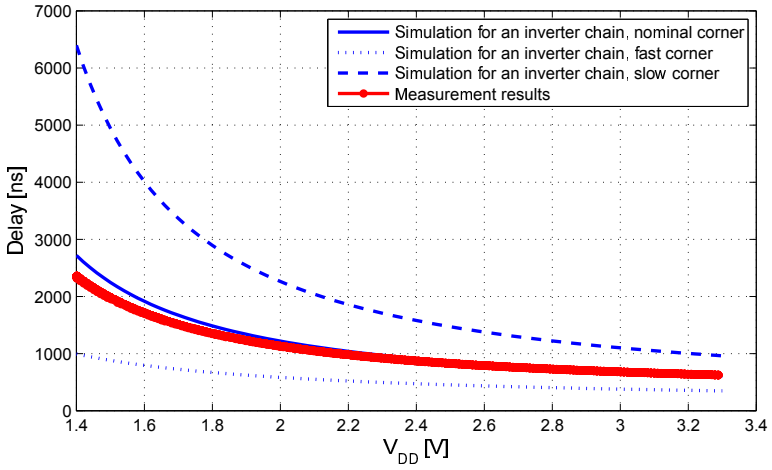
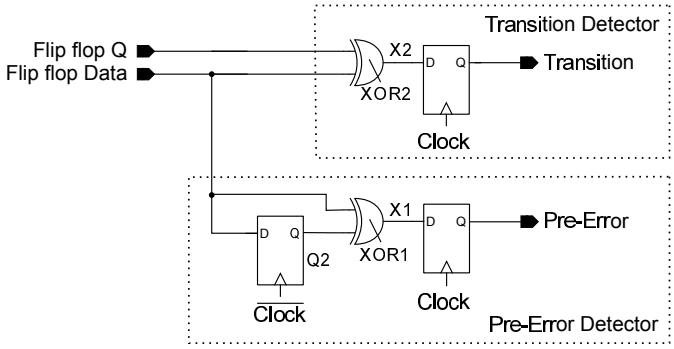


Figure 6.2.: Sensitivity of the digital logic in the digital front-end of NMS to variations, evaluated by simulations and measurement data.

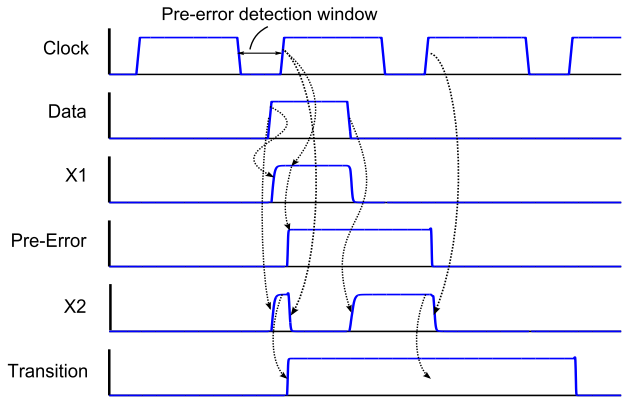
slack in paths equipped with monitors drops below a certain threshold. The number of pre-errors in a certain time interval, i.e. the pre-error rate indicates the circuit delay. In other words, the intermediate module averages the number of pre-errors over a time span. For zero pre-errors, the voltage is reduced. On the contrary, an increased pre-error rate results in an increase in the supply voltage, as it indicates a reduced circuit speed.

6.2.1. Integrated Monitor

The in situ delay monitors are able to distinguish between relaxed and critical operation of the circuit. To realize the pre-error flip-flop, extra circuitry is added to the flip-flop, as shown in Fig. 6.3a [90]. The pre-error detection window is determined by the duty-cycle of the clock signal e.g. the low phase of the clock for positive edge triggered flip-flop. A flip-flop with inverted clock as a clock input is added to the regular flip-flop. In this structure, to avoid the risk of timing errors by extreme scaling of the supply voltage in low activity phases, a transition detector is exploited. The transition detector monitors all data transitions, either relaxed or critical, to distinguish between active and inactive clock cycles. For the transition detector, the inputs of an XOR gate are the data signal and the output Q of the regular flip-flop. In case of a data transition, the input signal Data will differ from its value in the previous clock cycle, stored as Q. Hence, a transition signal is generated by the XOR gate. Fig. 6.3b shows the corresponding timing diagram for both pre-error and transition detector.
The sensitivity of the monitors to voltage reduction is analyzed and the accuracy of the window for pre-error detection is extracted through simulation. The monitors



(a)



(b)

Figure 6.3.: a) Schematic of the in situ delay monitor including the pre-error detector and the transition detector and b) the timing diagram of the monitor, pre-error detector and transition detector, in case of detecting a pre-error (data changing in the pre-error detection window)

are simulated considering the variation of the system clock. Figure 6.4 shows the sensitivity of the pre-error detection window for the 350nm technology, at a temperature of $Temp = 37\,^{\circ}\mathrm{C}$ (body temperature). Over all corners and for a supply voltage as low as $V_{DD} = 1.5\,\mathrm{V}$, the deviation of the detection window of the synthesized monitors is less than 0.1% of the operating clock periods, with a frequency of 3-8MHz.

6.2.2. Abstraction of the Monitor Data

Since the occurrence of one pre-error at the output of one single monitor in a certain module is sufficient for detecting a critical timing, the output signals of several pre-error flip-flops are fed into an OR tree. The OR tree combines the

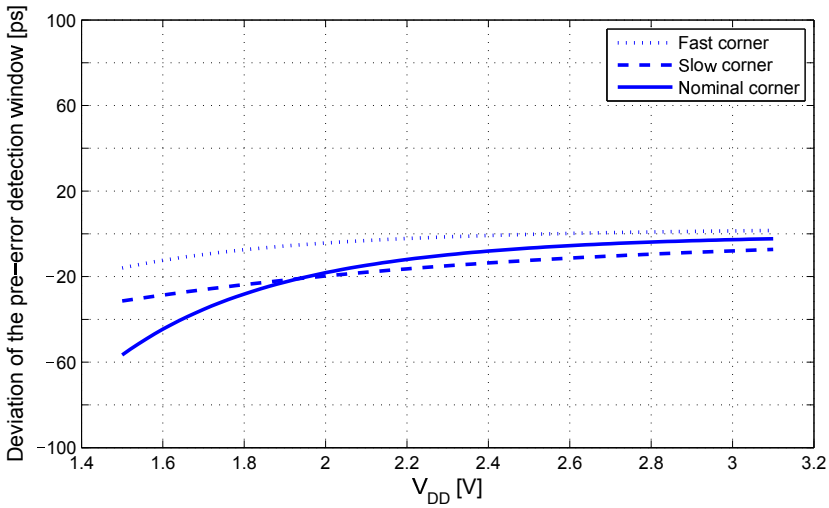


Figure 6.4.: Deviations of the monitor detection window over V_{DD} under process variation and voltage reduction

generated pre-errors and transitions as a single pre-error and transition, respectively. Processing of the monitor data is performed by the monitoring control unit, which counts the pre-errors in an observation interval comprising N active clock cycles, distinguished by the number of transitions. Thus, the pre-error rate is defined by averaging the number of pre-errors over all active clock cycles. The reason is that in digital circuits an activity rate of 100% is a rare condition. This means that data transitions at the end of combinatorial paths do not occur in every clock cycle. Therefore, during a fixed time interval some clock cycles have no data transitions. If regulating the supply voltage is based only on the number of occurred pre-error pulses in a fixed time interval the probability of detecting the pre-errors will reduce. Occurrence of no pre-errors does not necessarily mean a relaxed timing. Thus, ignoring the activity rate of the data signals might result in aggressive voltage reduction. An exceedingly low supply voltage increases the possibility of timing errors in following more active clock cycles. Thus, the observation interval includes N active clock cycles (a cycle in which a data transition occurs) instead of N fixed number of clock cycles.

6.2.3. Closed Loop Configuration

The control loop is shown in Fig. 6.5. The resulting pre-error count is translated to a two bit control word, CW_1, CW_0, which shows the reliability status of the circuit and can be used for the adaptation of the operating parameters. To elucidate the count of pre-errors in an observation interval, two limits for decision regarding the reliability status of the circuit are defined, $n_{limit\uparrow}$ and $n_{limit\downarrow}$. If the

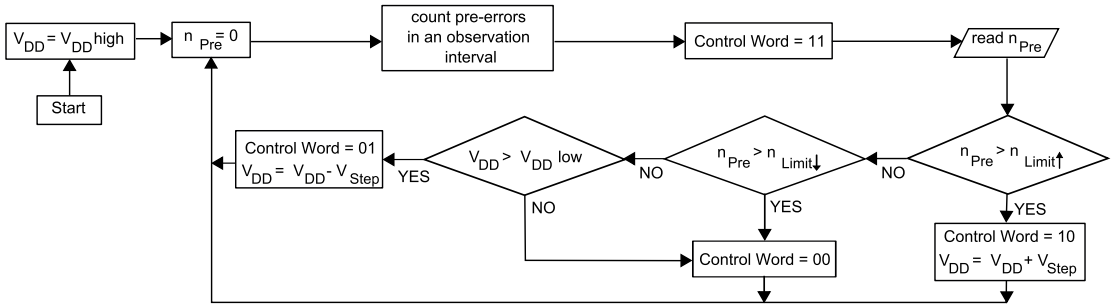


Figure 6.5.: Control loop of the in situ monitoring system for reliability monitoring and adaptive supply voltage regulation

resulting two bit control word is the binary number "11" a flag is assigned. The flag indicates that an observation interval is finished and the count of pre-errors is compared to the aforementioned limits. This means that in the next clock cycle the output control word shows the status of the timing. If the count of pre-errors, n_{pre}, is above the upper threshold of $n_{limit\uparrow}$, timing is critical and the control word equals "10". Therefore, the voltage has to be increased to increase the speed of the circuit and relax the timing. If n_{pre} is under the lower threshold of $n_{limit\downarrow}$, the timing is relaxed, the control word equals "01". Thus, the circuit is considered as reliable and the voltage can be decreased to a lower level. If n_{pre} is between $n_{limit\downarrow}$ and $n_{limit\uparrow}$ no action is required, i.e. the voltage is maintained, an idle state is assigned and the control word equals "00". As an example a conservative setting is used in which the voltage is reduced only if no per-errors occur. Therefore, observation interval is set to $N = 1024$ clock cycles, $n_{limit\downarrow} = 1$ and $n_{limit\uparrow} = 20$. Therefore, the voltage only decreases if no pre-errors occur.

6.2.4. Realizing the Detection Window

For high speed circuitry such as the clock divider which has the maximum frequency of the system, it is desirable to optimize the operations per cycle and thus have the minimum number of logic stages before the flip-flops. Therefore, a considerable part of the high speed clock period can be consumed for synchronization. Since the duty cycle of the clock signal is exploited as the detection window for the pre-error monitors and the monitoring takes place for both positive and negative edge triggered flip-flops, two monitoring clocks are required for pre-error detection. The on chip clock divider is configurable for different duty cycles for the monitoring clock. The duty cycle can be chosen between duty cycles of 1/3 and 2/3.
Figure 6.6 shows the clock pattern generator which is fed by an input high fre-

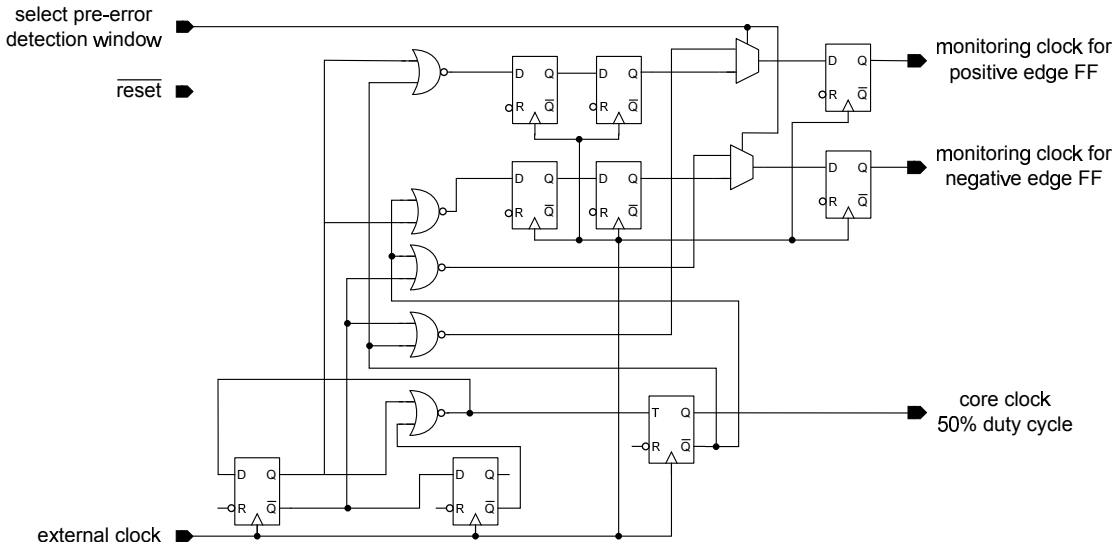


Figure 6.6.: On chip clock divider configurable for different duty cycles for the monitoring clock

quency clock. The clock generator operates error-free with supply voltages down to 1.1V in the slow corner. Therefore, the supply voltage of the system cannot be reduced beyond this limit. Moreover, for a target frequency, the entire monitoring circuitry needs to be evaluated to ensure the operation of the monitoring system itself under all possible supply voltages. This determines another limit for the minimum supply voltage of the circuit. For a core clock frequency of 4MHz, this limit is evaluated as $V_{DD,low} = 1.5$V in the slow corner and 0.9V in the nominal corner.

6.2.5. Monitor Placement

The in situ delay monitors should provide sufficiently accurate data while consuming small area and power. To optimally place monitors in the circuit, critical paths are identified. To optimally place the monitors within the core logic the simplified approach of section 4.2.1 is used. In the first step, to find the most critical paths a timing report was performed for the test chip. Here, only the critical paths of the high performance core clock domain are regarded. Out of the 250 most critical paths, the 16 most critical ones have been equipped with monitors. Moreover, 16 paths with the highest toggling activity are identified by the coverage analysis. The reason is to be able to quickly increase the supply voltage in case of occurrence of too many pre-errors. The resulting area overhead of the monitoring system integrated in the digital ASIC is only 2.9%.
The monitors are implemented in Verilog and synthesized together with the func-

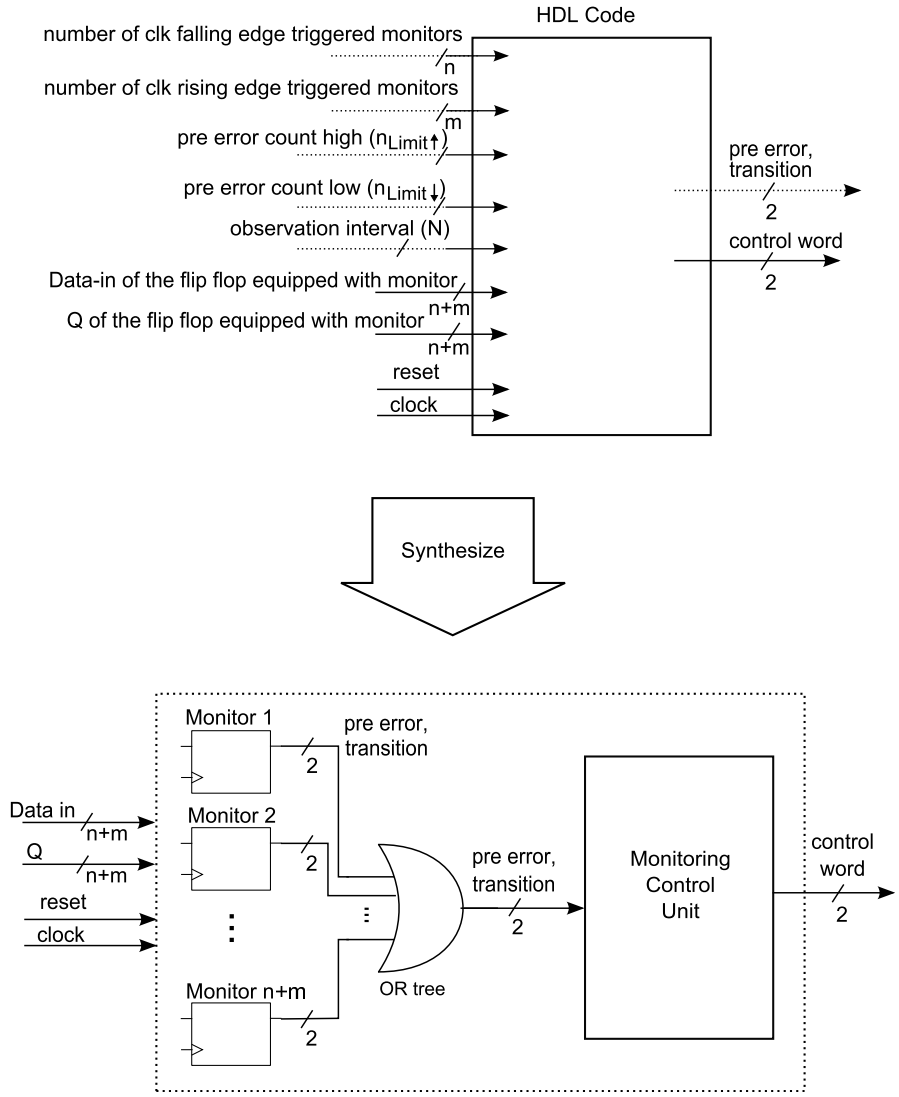


Figure 6.7.: Monitoring system comprising the in situ monitors and the monitoring control unit is implemented in Verilog and synthesized in a target technology.

tional logic. In the Verilog code, the number of required monitors, whether they are clocked by the negative or positive edge of the clock and the module in which the monitors are to be included are defined. Fig. 6.7 shows the monitoring system, with the in situ monitors and the monitoring control unit which are described in Verilog and synthesized in a target technology (here 350 nm CMOS technology) using the RTL-Compiler.
The pre-error flip flops can also be used to indicate the reliability status of the signals entering a certain module. Therefore, they can also be integrated at the

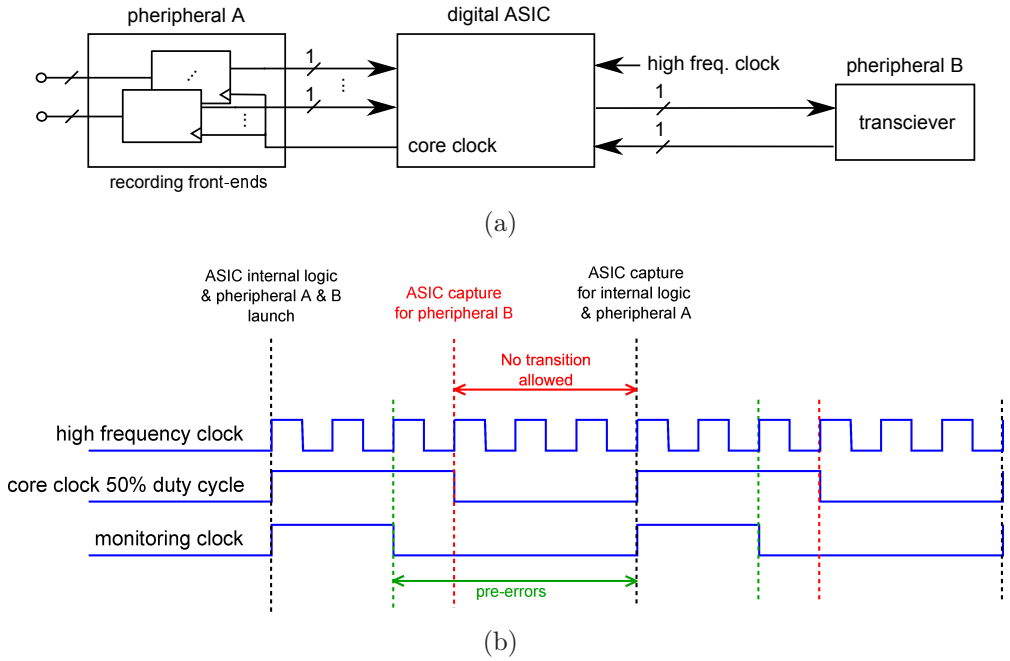
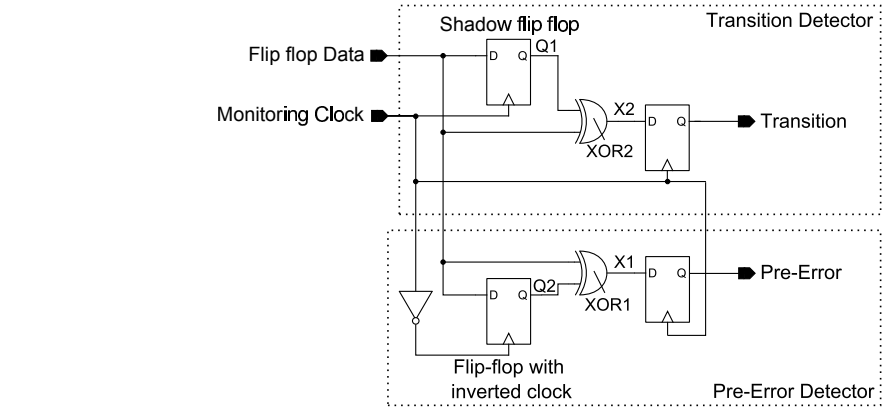
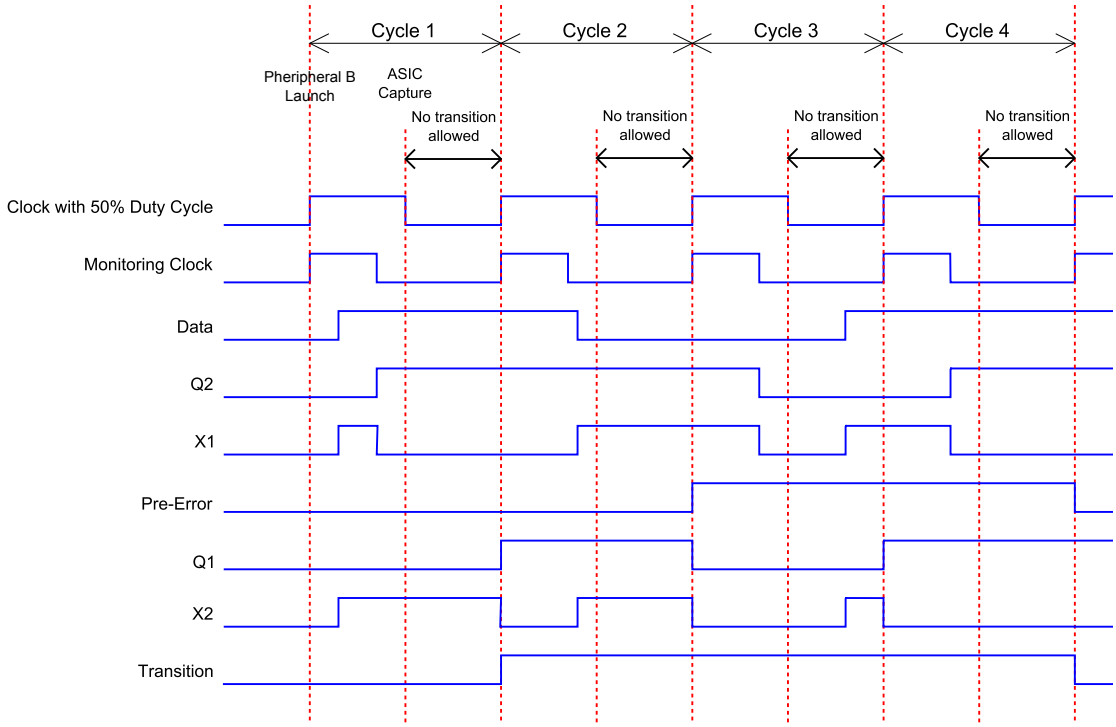


Figure 6.8.: Connections of the Neuro digital ASIC to the peripherals ([100]) and the monitoring clock for the peripherals

input pads of the digital front end module for voltage scaling of the whole system. In Fig. 6.8a the connection of the digital Neuro-ASIC to the peripherals is shown. Peripheral A writes data to the output on the rising edge of the core clock. Nevertheless, this domain is different to the internal core clock domain, as the clock signal is fed out through a PAD-cell resulting in an additional skew compared to the digital ASIC clock. The number of the monitors used for every domain depends on the number of critical paths to monitor. In Fig. 6.8b the monitoring clock for the peripherals is shown. Since both modules write on the rising clock edge with a small delay, the same monitor clock is used for both modules. Monitors used for the outputs of the peripheral A are the same as the monitor shown in Fig. 6.3a. At the inputs of the ASIC from peripheral A, 8 monitors are inserted. To monitor the peripheral B, the monitor structure is modified and a shadow flip-flop is added, as in Fig. 6.9a. The timing diagram is shown in Fig. 6.9b. Here, the data is launched by the positive clock edge and captured by the next negative clock edge. In addition to data transitions violating the setup time for the negative edge triggered flip-flop, transitions occurring after the negative clock edge result in an error. Therefore, the detection window begins after the positive clock edge and before the negative clock edge and continues until the next positive clock edge. Thus, for data transitions occurring during the detection window, whether the data transition is before or after the clock falling edge a pre-error is assigned



(a)



(b)

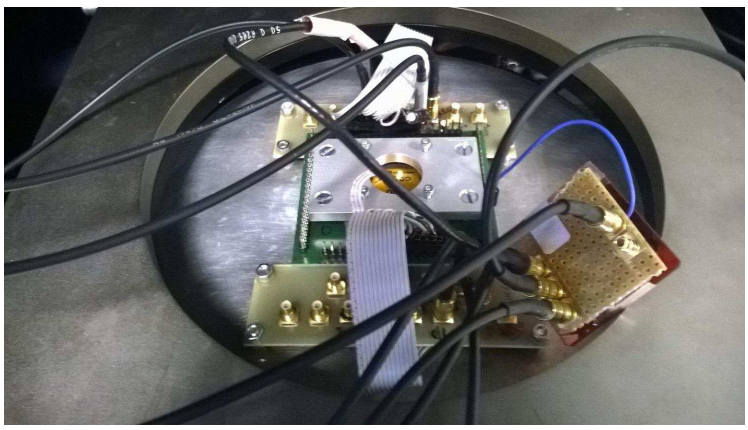
Figure 6.9.: a) The schematic of the monitor modified for peripheral B and b) the timing diagram of the corresponding monitor. When a data transition happens during the detection window, i.e. when the monitoring clock is low, a pre-error is generated.

(both critical and erroneous signals). At the input of the ASIC from peripheral B one monitor is inserted.

6.3. Experimental Results

The efficiency of the in situ monitoring approach for reliable and power efficient digital design of the NMS is evaluated by experiments on the fabricated chips. The measurement setup evaluates the effect of voltage reduction on the test chip. An on-chip ring oscillator is implemented. To observe the performance reduction by lowering the supply voltage, the frequency of the ring oscillator is measured. Moreover, the frequency degradation of the ring oscillator after stress indicates the effect of aging mechanisms on the digital circuitry. In the measurement setup, stress cycles are included to estimate the effect of degradation mechanisms over the lifetime of the implant. This is due to the fact that, the implants have very strict reliability requirements and need to operate error-free for a long lifetime (up to decades, e.g. 30 years). By adapting the supply voltage to minimal values, the performance degradation of the circuit is reduced.
To perform the stress and measurement cycles, the chips undergo burn-in phases of 1, 2 and 4 hours and so on with an accelerated supply voltage of 3.8-4.5V and a temperature of $Temp = 125\,^{\circ}\mathrm{C}$. After each stress cycle there is a measurement cycle evaluating the minimal reliable supply voltage. Figure 6.10 shows the measurement setup and the connections to the test chip.
To precisely observe the resulting control words which are provided to the voltage regulator, first an open loop measurement performed. Thereby, during each measurement cycle the supply voltage is set to the nominal value and afterward is slowly decreased. Figure 6.11 shows the chip temperature, the resulting on chip supply voltage and the ring oscillator frequency against time for the corresponding open loop configuration. As shown in Fig. 6.11, the control word switches from the relaxed state of CW_1, CW_0 ="01" to the critical state of "10" within the measurement cycles by lowering the supply voltage. The crossover voltage has an increase of 25mV through 3 burn-in phases. The control word switches from "10" to "01" by increasing the supply voltage and entering the stress phase. Immediately after each stress phase there is an interval in which the supply voltage is decreased, but the temperature is yet higher than the nominal value of $Temp = 37\,^{\circ}\mathrm{C}$. The reason is that, during this interval the temperature decreases slower than the supply voltage and finally reaches the nominal value of $Temp = 37\,^{\circ}\mathrm{C}$. Thus, in this interval the ring oscillator frequency is increased as the supply voltage decreases faster than the temperature. An idle time is assigned to reach the low measurement temperature and supply voltage.
The effect of stress on monitors is negligible due to the low supply voltage. However, a slightly increased setup time of the pre-error flip-flop is partly compensated by an increased delay of the inverter for the clock input. For applications that experience a high performance degradation by aging mechanisms it is also possible to design aging resistant monitors as shown in section 5.1.3. Figure 6.12 shows the frequency at constant supply voltage of 1.5 V against the time of burn-in.

(a)



(b)

Figure 6.10.: a) Measurement setup b) connections to the test chip

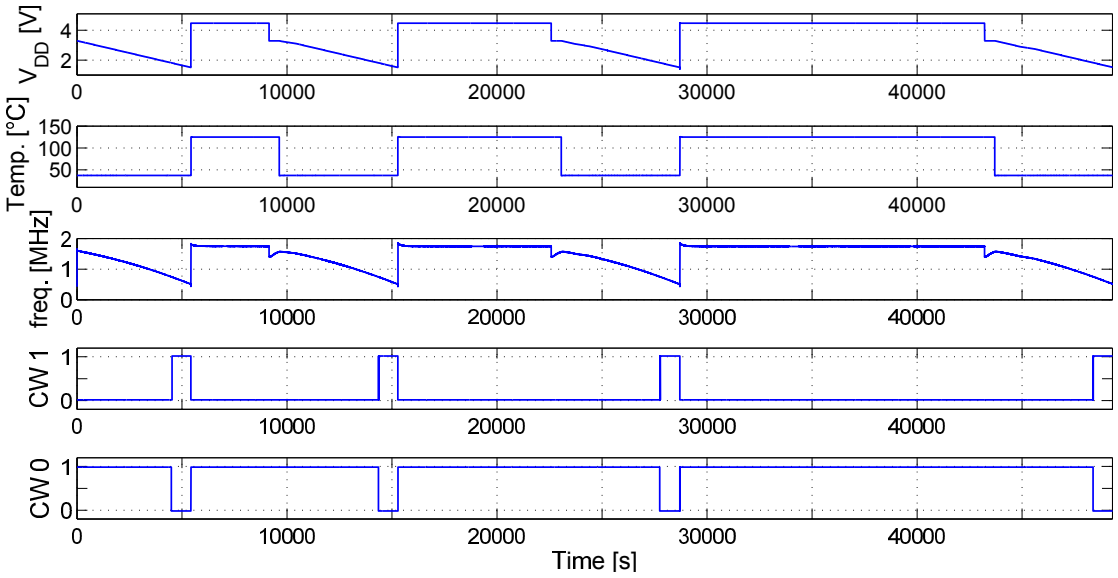


Figure 6.11.: On chip voltage, temperature and ring oscillator frequency plus the resulting control words against time for measurement and burn-in/stress phases with open loop voltage regulation.

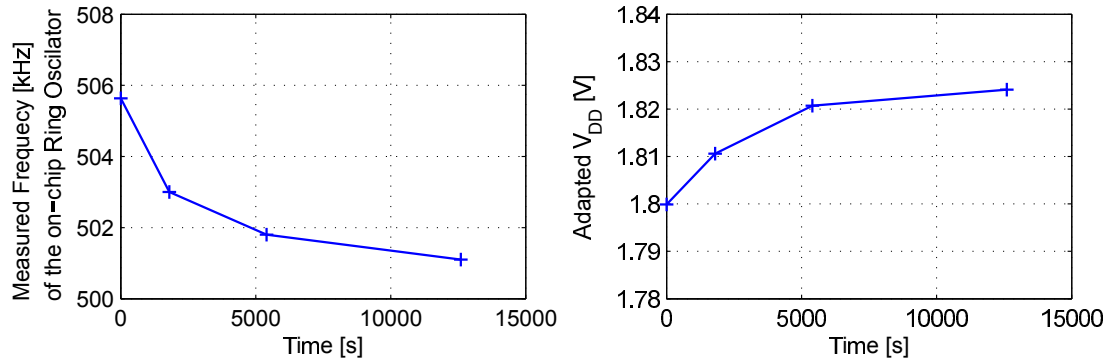


Figure 6.12.: a) Frequency of the on-chip ring oscillator after several stress phases measured at a constant supply voltage of 1.5 V. b) Adapted supply voltages against stress time for a core clock frequency of 4MHz, and detection window of $0.66 \cdot T_{clk}$

In Fig. 6.12a, a frequency decay of 4 kHz can be seen between 3 stress phases. Moreover, in Fig. 6.12b the supply voltage in which control words switch is also depicted.
In the next step, the supply voltage is regulated by a closed loop configuration. The voltage regulation is performed off-chip by evaluating the resulting control words. Based on the resulting control words for a certain monitoring setup, the

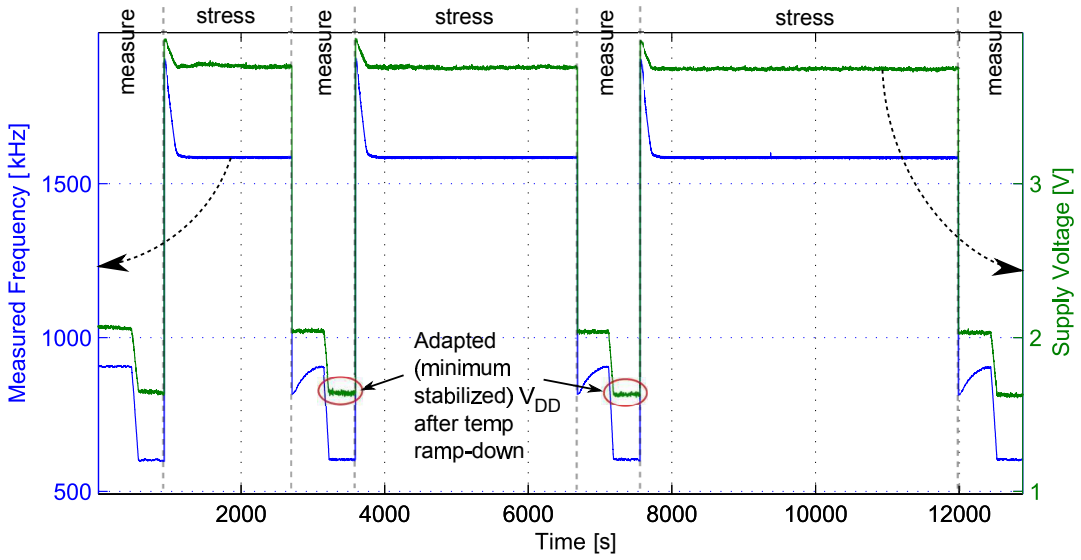


Figure 6.13.: On chip voltage and the resulting ring oscillator frequency against measurement time for measurement and burn in/stress phases, with closed loop voltage regulation during measurement phases

voltage is regulated and a minimum operating supply voltage is assigned. Thus, by assigning the $n_{limit\downarrow}$, $n_{limit\uparrow}$ and N (number of active clock cycles as observation interval) for defined reliability criteria and a certain clock frequency, a corresponding power reduction is achieved. Figure 6.13 shows the resulting stress-measurement cycle and the adapted operating supply voltage. The adapted operating supply voltage is strongly dependent on the core clock frequency and the remaining timing slack. In Fig. 6.13 a core clock frequency of 3.33MHz is assigned and a adapted operating supply voltage of 1.6 V for a detection window of $0.66 \cdot T_{clk}$ is achieved. Figure 6.14 shows the adapted supply voltage determined by AVS which is de-

clock freq.	guard-banding V_{DD}	AVS V_{DD}	guard-banding power	AVS power	power saving
6MHz	2.00V	1.56V	2.70mW	1.60mW	41%
8MHz	2.26V	1.80V	4.64mW	2.85mW	39%

Table 6.1.: Measured power savings by the adaptive voltage scaling utilizing the monitoring approach compared to guard-banding approach for the fresh circuit

pendent on clock frequency and monitoring settings. Moreover, in this figure the state of the art guard-banded supply voltage for each clock frequency is depicted,

clock freq.	guard-banding V_{DD}	AVS V_{DD}	guard-banding power	AVS power	power saving
6MHz	2.03V	1.60V	2.77mW	1.68mW	40%
8MHz	2.30V	1.86V	4.87mW	3.06mW	37%

Table 6.2.: Measured power savings by the adaptive voltage scaling utilizing the monitoring approach compared to guard-banding approach assuming a worst case performance degradation of 5%

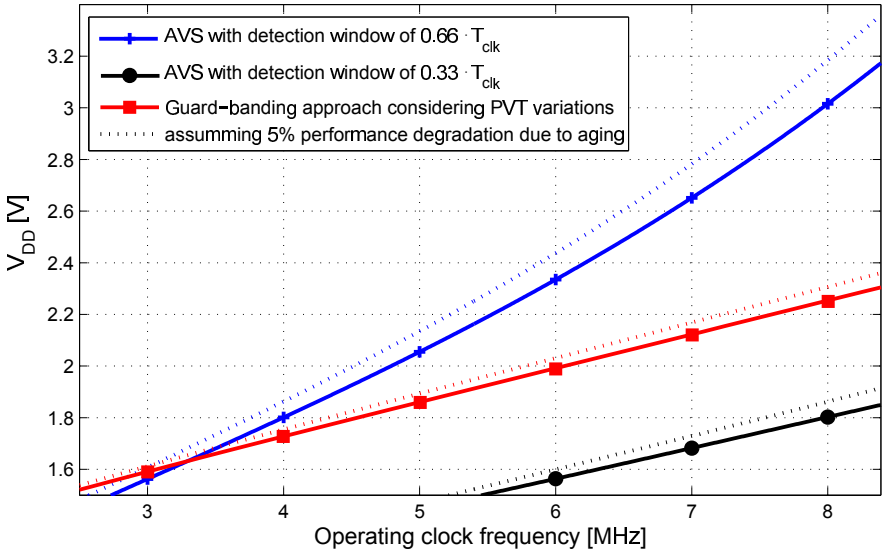


Figure 6.14.: Dependence of adapted and guard-banded supply voltages on operating clock frequency for a certain monitoring setup.

considering the worst case guard banding approach (voltage drop of 10%, slow process, body temperature of $Temp = 37\,°C$). For the AVS scheme two monitoring clock duty factors of 33% and 66% are shown with detection windows of $0.66 \cdot T_{clk}$ and $0.33 \cdot T_{clk}$, respectively. In Fig. 6.14 the dashed lines consider a performance degradation of 5% due to aging. Nevertheless, this is only a hypothetical value for the 350nm technology.
At a clock frequency of 8 MHz, a supply voltage of 2.26 V is assigned for the guard-banding approach. Assuming a hypothetical 5% aging during the lifetime of the circuit, the guard-banded supply voltage increases to 2.30 V. Applying the AVS with a detection window of $0.33 \cdot T_{clk}$ reduces the supply voltage of the new and aged circuit to 1.80 V and 1.86 V. The measurements show the quadratic behavior of power consumption in dependence on supply voltage. While the circuit still operates error-free and reliable, power savings of 37% and 39% are achieved with

and without considering the aging, respectively (see Tables 6.1 and 6.2).

6.4. Summary

In applications with strict reliability requirements, the presented in situ monitoring approach enables dynamic adaptation of operating parameters such as supply voltage and thus decreases the power consumption of the system. By minimizing the power consumption of the NMS and yet maintaining the required reliability, the operating temperature of the implants is controlled. Thus, the potential health risks to the patients are reduced. Moreover, by assigning the lowest possible supply voltage while maintaining the required reliability, device aging is reduced and the lifetime of the circuits is prolonged. In case that errors occur during the operation, the supply voltage can be adapted to a higher value, which ensures continuous correct operation of the NMS system for many years. The quantitative evaluations on simulation and measurement results support the applicability of the reliability monitoring methodology for the neural measurement system.

7. Evaluation of the Monitoring System in Automotive Applications

Nowadays, more and more electronic devices are being used. In safety critical applications, electronic devices must satisfy certain reliability specifications, which are defined by the lifetime requirements. For example, modern cars exploit more than 80 electronic control units (ECUs) [73]. These semiconductor devices enable efficient, safe, comfortable and better performing operation of the vehicles [104, 71]. However, devices used in vehicles have to withstand wide range of electrical transients, strong thermal and mechanical stresses [105]. Moreover, an increased functionality in modern high-end cars is demanded. Thus, the electronics in the car is further increasing [71, 106] and new standards are introduced to comply with functional safety [107, 108]. Therefore, satisfying high reliability requirements for such applications is a crucial design goal.
Traditionally, for safety critical applications during the design phase extreme corner cases were analyzed. However, regardless of the circuit's application, since the 40nm technology node, reliability assessment is a must during the design phase. Afterward, during the qualification phase, reliability tests simulate the actual lifetime stress for the electronic component. This chapter proposes to track the reliability status of a circuit with high reliability requirements during the lifetime. Thus, in the monitoring approach presented here, in situ monitors are used to check the reliability of a circuit under test. All the circuits presented in this chapter are designed and fabricated in a 40nm technology. In the proposed monitoring approach in situ timing monitoring of the digital logic can be used for reliability diagnosis. Moreover, the monitoring approach enables to take the necessary countermeasures such as adaptation of operating parameters.

7.1. Circuit under Test

To analyze the effect of stress conditions on the circuit and to track the aging of the devices in 40nm technology, the core logic components are exploited as the circuit under test (CUT). Thus, CUT includes 15 logic paths equipped with the monitors developed in this work. At the end of the paths the custom designed

monitors extract the remaining slack of the path. The paths under test include the critical path of an arithmetic circuit and combinations of the standard library elements. Therefore, besides the logic chains in the circuit under test the critical path of a synthesized circuit is extracted and characterized by the in situ monitors. Table 7.1 shows the type of the elements used in the paths under test. It should

path number (binary)	path elements	monitor type	distance to next module
0000	not used in the circuit under test	-	-
0001	critical path of an arithmetic circuit	3	b
0010	inverter chain	3	a
0011	inverter chain	3	b
0100	2-input NOR chain	3	a
0101	2-input NOR chain	3	b
0110	3-input NOR chain	3	a
0111	3-input NOR chain	3	b
1000	3-input NAND chain	3	a
1001	3-input NAND chain	3	b
1010	short 3-input NOR chain	3	a
1011	short 3-input NOR chain	3	b
1100	3-input NOR chain	2	a
1101	3-input NOR chain	1	b
1110	20% inverter, 30% 2-input NOR, 50% 3-input NOR	1	a
1111	40% inverter, 60% 2-input NOR	1	b

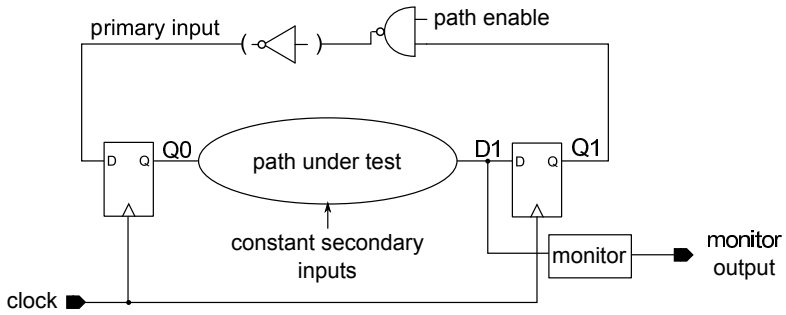
Table 7.1.: Type of elements used in the path under test. Monitor types 1, 2 and 3 are shown in Figures 5.8, 5.11 and 5.15, respectively. Position “a” is very close to the next processing module and position “b” has a certain distance to the next processing modules.

be noted that even if some paths in Table 7.1 are similar, they are equipped with different monitors and have different distances to the next processing module, position “a” is very close to the next processing module and position “b” has a certain distance to the next processing modules. This enables to evaluate the accuracy of the monitors as well as the effect of distortion by long wires on the pulses generated by the monitors.
To simplify the control of the test patterns applied to the paths under test, a clocked ring oscillator is designed which includes the path under test, as shown in Fig 7.1a. The corresponding timing diagram is shown in Fig 7.1b. In the clocked ring oscillator circuit shown in Fig 7.1a, the path is stimulated through an oscillating signal synchronized to the core clock. Thus, no external stimuli are

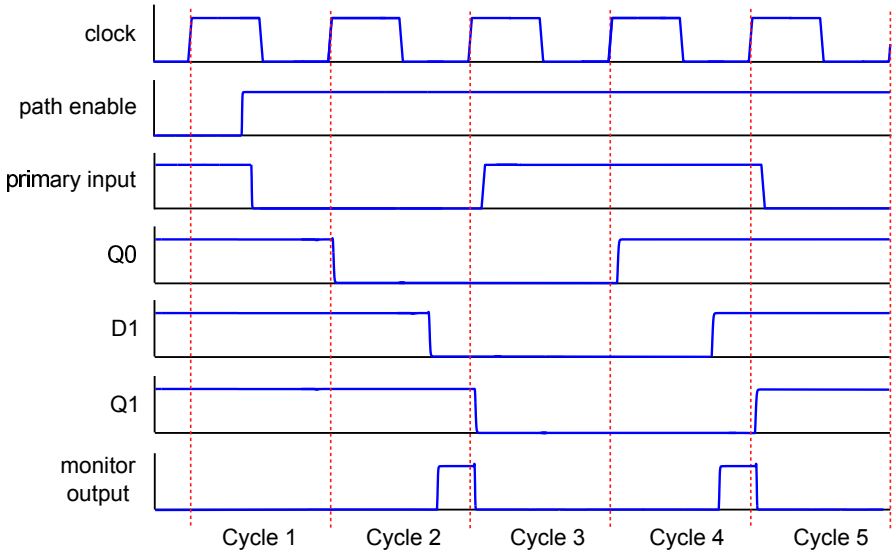
required to be applied to the path under test. When the path-enable is deactivated (logic "0"), the primary input is logic "1" (or "0"). When the path-enable signal transitions to logic "1" (cycle 1) the primary input transitions to logic "0" (or "1") at the same clock cycle. Thus, $Q0$ is updated to the new value of the primary input by the clock triggering edge (cycle 2). The value change of $Q0$ results in a data transition at the node $D1$, with the delay of the path under test. Afterward, $Q1$ is updated by this value at the next triggering edge of the clock signal (cycle 3). As the path-enable signal is logic "1" the NAND gate operates as an inverter and the primary input is inverted at cycle 3. Thus, an oscillation occurs which is synchronized to the internal core logic and has a period of 4 times the clock period. This is similar to a divide by $n = 4$ operation. This means that a falling/rising data transition at node $D1$ happens every 4^{th} clock cycle. By inserting extra flip-flops after $Q1$, n can be changed to the desired value. Adding a flip-flop stage together with a dummy logic (buffers) will change the occurrence of monitor outputs to every 3^{rd} clock cycle. The dummy logic ensures that the hold time constraint of the flip-flops is not violated. In the CUT, 3 flip-flops are used to relax the timing constraint for the control signal for the monitors (Monitor-Enable) and the implementation of the TDC.

7.2. Monitoring System

The developed in situ monitoring approach is implemented to evaluate the feasibility and the benefits of the reliability management system utilizing the run time monitors. As mentioned before, the monitoring system is applied to the digital logic (ASIC, i.e. the CUTs presented in the previous section). The degradation level is monitored by in situ timing measurement of the paths under test, where several monitor structures are implemented. Monitors provide the remaining timing slack as a pulse to a 5-bit delay line time to digital converter. The path under test is chosen by a multiplexer (MUX) tree and the TDC provides the remaining timing slack of the corresponding paths under test as a digital code. The calibration circuitry of the TDC is implemented on chip. Moreover, the on chip control circuitry enables different scenarios for qualification of the circuit under test. The possible scenarios include DC or AC stress conditions as well as measurement cycles. Figure 7.2 shows the top level view of the monitoring system and the paths under test. According to the SPICE simulations the total power consumption of the test chip for a measurement cycle is 3.5mW.
The monitored circuitry consists of two similar CUTs. The upper one shown in Fig. 7.2 is equipped with precise slack monitors (section 5.1.3), as discussed in Table 7.1. The lower CUT includes the same paths but equipped with 2-bit monitors (section 5.1.2). Figure 7.3 shows the layout of the developed circuitry in the test chip.



(a)



(b)

Figure 7.1.: a) Path under test, in situ monitor, launch and capture flip-flops and extra circuitry for control purpose. When the path-enable signal is activated the path oscillates synchronized by the clock signal. b) The timing diagram of the paths under test

7.2.1. Input/Output Interfaces

The input interface (Fig. 7.4) uploads the control signals serially into the test chip. It also decodes the control signals before providing them to the target modules. As the decoding is performed synchronized to the slow external clock, timing violations in the decoding circuitry do not occur. The control signals include the select word for the tunable ring oscillator, select bits for the TDC calibration circuitry, path-select signal for selecting the path under test by the multiplexer tree, and the enable bit for the monitors. When the control bits are uploaded, the ring-oscillator-enable signal disables the shift registers. After one triggering

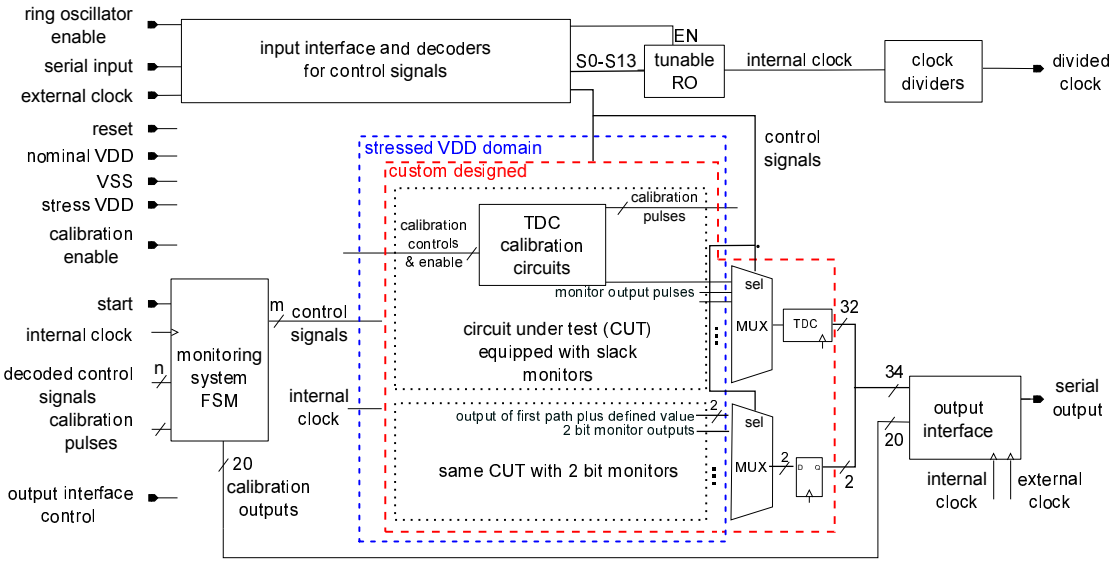


Figure 7.2.: Top level view of the implemented circuits on the test chip

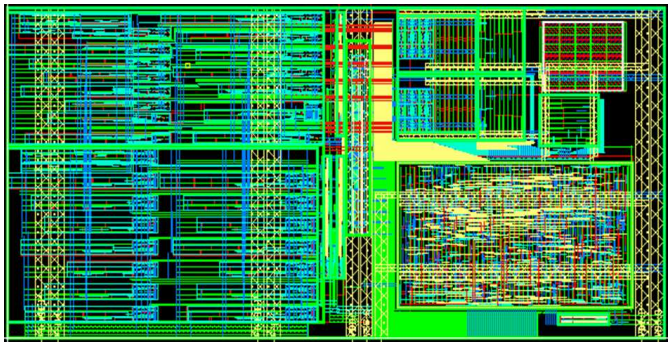


Figure 7.3.: Layout of the implemented circuits on the test chip

edge of the external clock signal, the ring-oscillator-enable signal enables the ring oscillator. Therefore, the select bits of the ring oscillator are stabilized before the ring-oscillator-enable-signal is triggered.
The measurement results are loaded into the output interface. Afterward, the output interface serially shifts out the results synchronized to the external clock. Here, based on the value of the output-interface-control (Fig. 7.2) the resulting data is either written synchronized to the internal clock or is read out by the external clock. All signals passing through the clock domains are synchronized by additional flip-flops to avoid timing violations in the new clock domain.

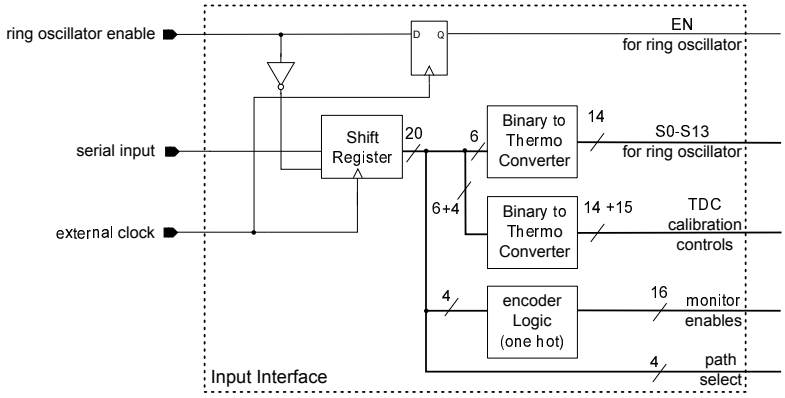


Figure 7.4.: Structure of the input interface

7.2.2. Clock Generator

In order to exploit different run time frequencies and thus different remaining timing slacks a tunable ring oscillator is designed and implemented (see Fig. 7.5). The tunability of the ring oscillator increases the testability of the system. The generated clock frequency is divided and can be observed by an output pad, as shown in Fig. 7.2. The control signals of the ring oscillator are applied by the input interface. The corresponding control circuitry of the input interface for the clock signal consists of shift registers and binary to semi-thermometer code. The uploaded and the encoded control word is directly applied to the select bits of the MUXes shown in Fig. 7.5. The reason for 64 tunable states is that the RO needs a high tunable range to provide different frequencies and scenarios for the slack monitoring in different parts of the CUT. In the design shown in Fig. 7.5, the control word is separated in two parts, a 3-bit binary most significant bits (MSB) and a 3-bit binary least significant bits (LSB). The encoded binary word in the input interface results into two thermometer codes for the MSB and LSB parts. When all select bits are deactivated (logic “0”) the ring oscillator delivers its highest frequency. Activating each bit in the MSB adds one of the fixed delay lines to the ring oscillator and increases the clock period, resulting in a lower frequency at the output. Activating each select input bit in the LSB increases the input period by approximately 4 times the delay of one inverter and decreases the resulting frequency accordingly. Figure 7.6 shows the post layout resulting clock period in dependence of the applied control word.

7.2.3. Selecting the Paths Under Test

The MUX tree selects the pulse representing the remaining slack of one specific path under test, determined by the path-select signals. Figure 7.7 shows the Monte Carlo simulations for local variations with 1000 runs for a tree of multi-

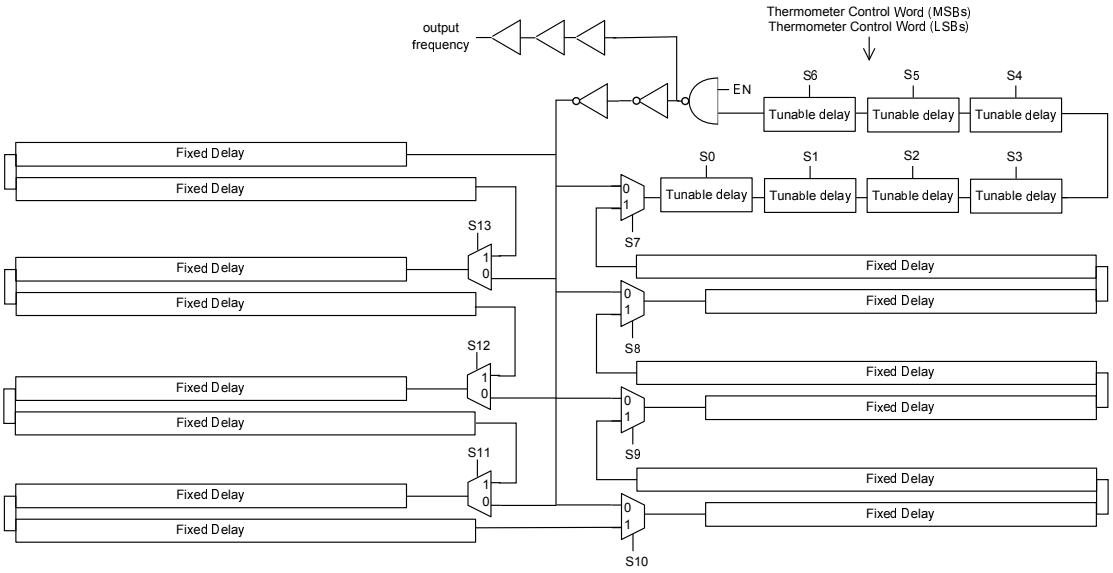


Figure 7.5.: Ring oscillator structure used as the internal clock generator

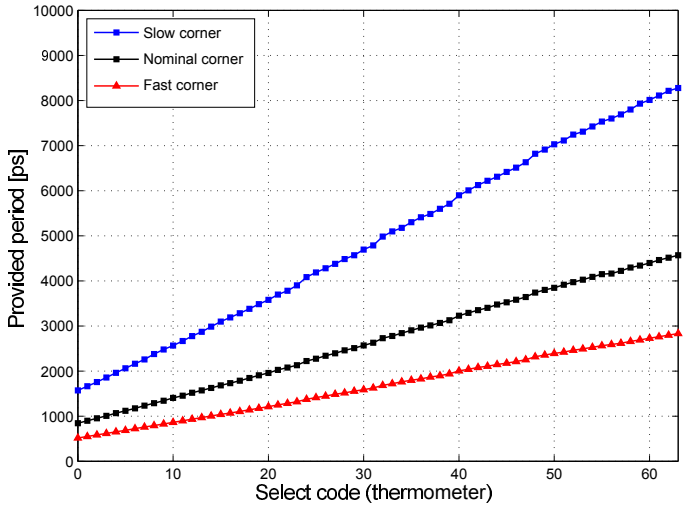


Figure 7.6.: Ring oscillator post layout simulated periods in different corner cases in dependence of the input control word

plexers, starting from a 2 to 1 and finally a 64 to 1 MUX. The MUXes designed for the standard library are normally optimized in terms of area and speed and not symmetry. Therefore, the MUX tree is custom designed to increase the symmetry regarding the inputs. The post layout maximum deviation of the inputs in the nominal case is 13ps. Figure 7.8a shows the possible design of the custom designed

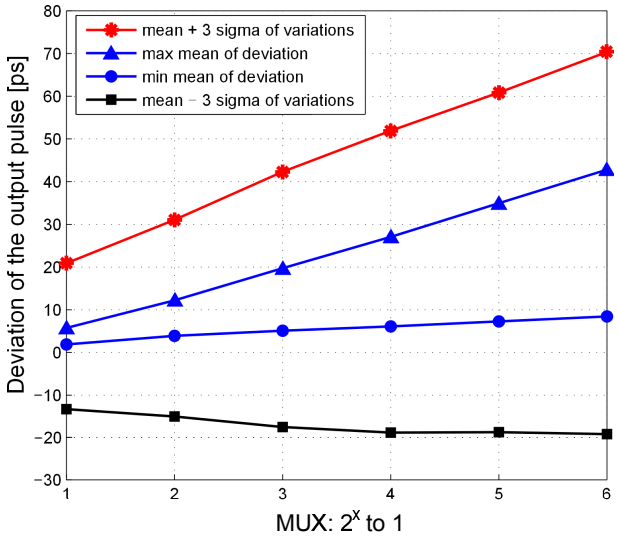


Figure 7.7.: Maximum output deviations from the input pulse for a MUX structure using standard library elements

MUX structure to increase the symmetry in correspondence to its inputs. Figure 7.8b shows the post layout simulation results for a 16 to 1 MUX tree including corner cases and local variations evaluated by the Monte Carlo simulations.

7.2.4. Integrated Monitors

Precise Slack monitors

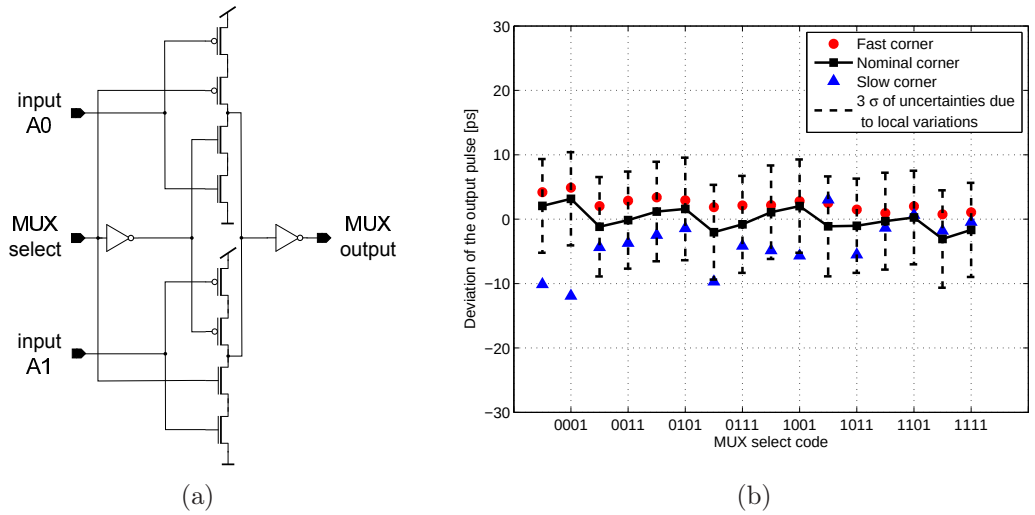
Three slack monitors which were discussed in section 5.1.3 are designed and fabricated. Figure 7.9 shows the post layout simulation results including the corner cases and local variations evaluated by the Monte Carlo simulations.

2-bit in situ TDC Monitor

In parallel to the approach with the precise slack monitors and the time to digital converter, same paths under test are equipped with 2-bit in situ TDC monitors (see section 5.1.2). Two structures, with delay elements (4 inverters) and without delay elements are designed. Figure 7.10 shows the results of this monitor when including 4 inverters as the delay elements in the monitors.

7.2.5. Slack Measurement by TDC

The operating frequency range is from 200MHz to 1GHz. The clock period is adjusted to generate different timing slacks. The delay degradation for the maximum



(a) (b)

Figure 7.8.: a) 2 to 1 custom designed MUX structure. The MUX tree is built by 4 stages of this structure, in total $8 + 4 + 2 + 1 = 15$. The layout of the whole structure is as symmetric as possible to minimize the deterioration for the pulses through different paths in the MUX structure. b) Post layout simulation results of the custom designed MUX tree (16 to 1) including the corner cases and the Monte Carlo simulations considering local variations

frequency determines the resolution. A 5-bit delay line TDC with 31 thermometer code output bits is designed. The TDC captures one pulse and goes to the disable state (one-shot measurement). Figure 7.11 and Figure 7.12 show the post layout simulation results. It is necessary to avoid the aging of the TDC. Thus, in the test circuitry two voltage domains are assigned. One domain is for the paths under test and monitors with the accelerated supply voltage and another with the nominal supply voltage for the rest of the circuit. The TDC calibration pulse is generated on chip and applied to the TDC by the controlling signal. The calibration pulse is generated at the first input of the MUX tree and is applied to the TDC when the path-select is "0000". For the coarse tuning of the calibration pulse an approach similar to the ring oscillator is used, similar to Fig. 7.5.

TDC calibration circuitry

In order to characterize the implemented TDC, a reference input pulse extracted from an start and an stop event is required. However, two independent signal sources are not suitable as they suffer from uncorrelated jitter [98] and the start and stop signals need to be generated by the same source. Thus, a calibration unit is implemented, as shown in Fig. 7.13. The calibration-enable generates a

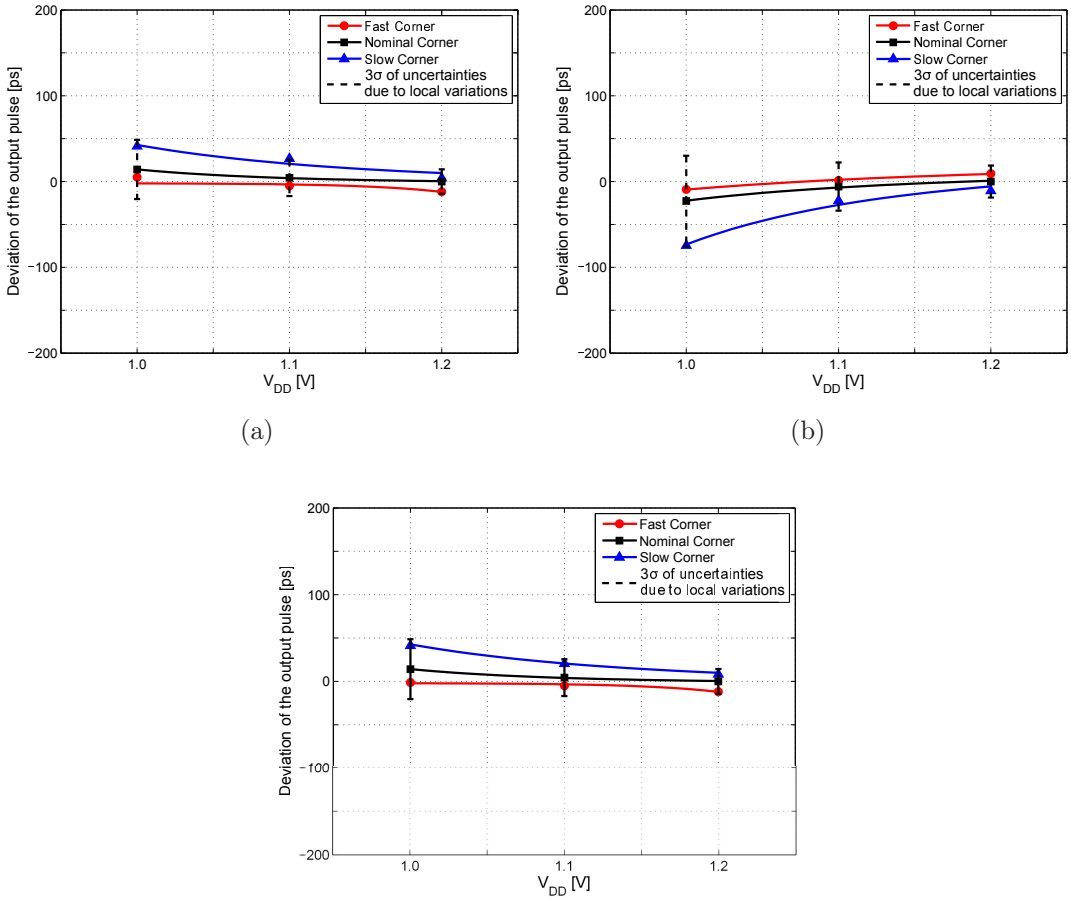
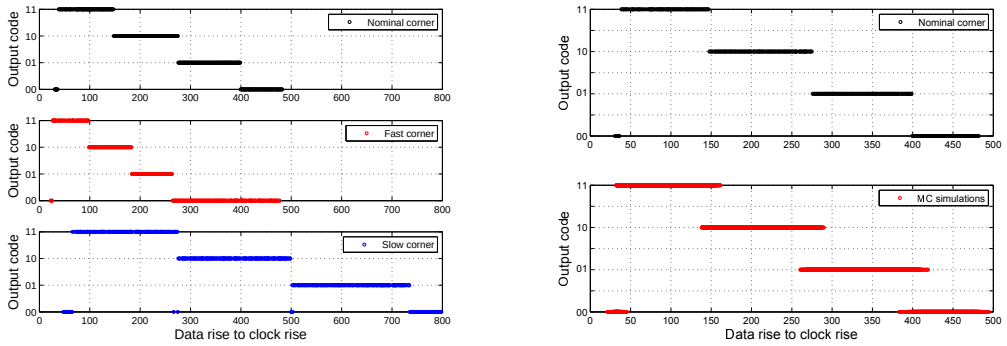


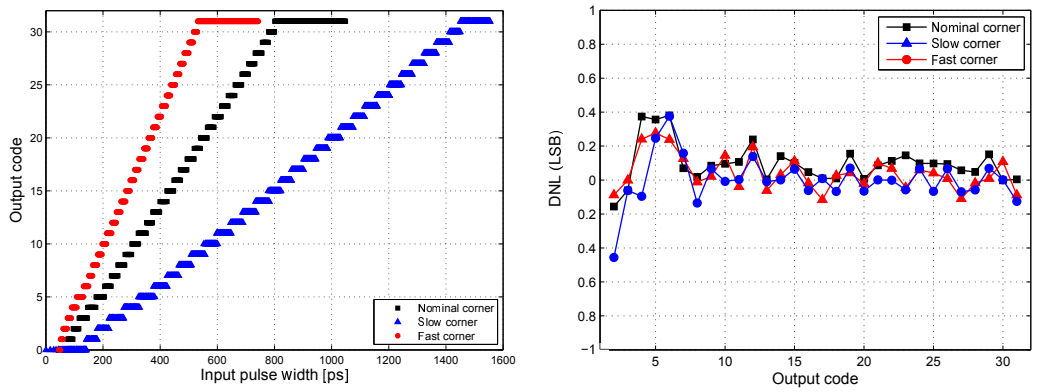
Figure 7.9.: Deviations of the output pulses for the precise slack generators, post layout simulations a) the low power latch slack generator (Fig. 5.8, type 1 in Table 7.1), b) the aging resistant dynamic slack monitor (Fig. 5.11, type 2 in Table 7.1) and c) the aging resistant static slack monitor (Fig. 5.15, type 3 in Table 7.1)

transition propagating through the start and the stop delay lines [98]. In the next step, a calibration pulse is generated and applied to the TDC.
Out of the single characterization pulse, a signal is sent into two independent delay chains: a fixed-delay chain of elements for the start and a tunable-delay chain of elements for the stop signal. The tuning code controls the amount by which the stop signal is delayed from the start signal. The start-stop delay can be tuned in two steps: a coarse and a fine tuning. The delay line that generates the calibration stop pulse is made out of three major parts. First, a series of buffers is inserted to reduce the resulting frequency as the clock of the stop counter. This is similar in



(a) (b)

Figure 7.10.: Post layout results of the output bit word for the 2-bit in situ TDC monitor, a) corner cases and b) Monte Carlo simulations considering local variations



(a) (b)

Figure 7.11.: Results of the corner analysis for a) the TDC full range b) the DNLs for each output word

both calibration start and stop delay lines. The second part is the coarse tuning delay part. The third part is the fine tuning circuitry. The start signal generation circuitry includes similar structure as the stop delay line but with fixed setting at the minimum delays.
The coarse tuning structure is similar to the ring oscillator structure shown in Fig. 7.5. Each multiplexed structure increases the start-stop difference by 29.6ps in the nominal case.
The aim of the fine tuning part is to generate sub-gate delays. This means that delays lower than the delay difference between the upper and lower path of the

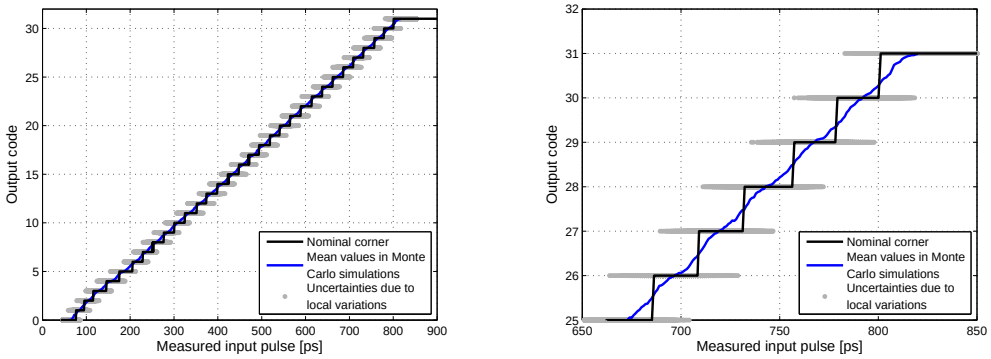


Figure 7.12.: Results of the Monte Carlo Simulations for a) the TDC full range b) the zoomed in uncertainties due to local variations

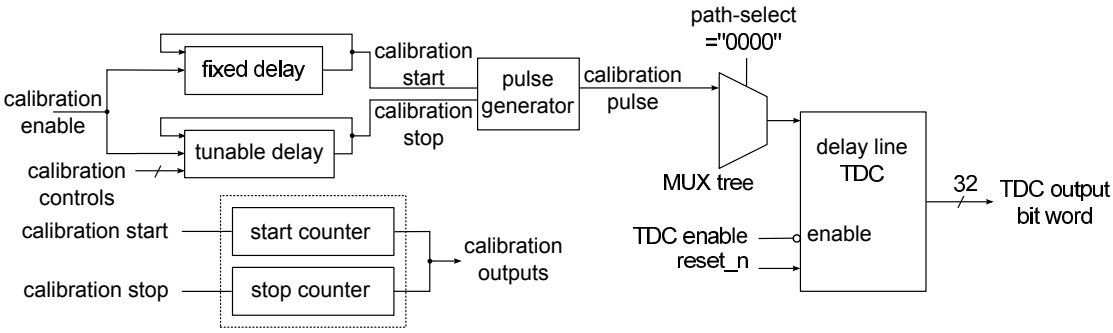


Figure 7.13.: Calibration circuitry of the TDC

multiplexer structure have to be generated. The circuitry that realizes the fine tuning is introduced after the coarse tuning part on the delay lines. The fine-tuning elements have to be controlled in such a way, that a difference in delays lower than the TDC resolution can be achieved between the start and the stop signal. Here, analog current-starved tunable-delay inverters or shunt-capacitor tunable-delay inverters can be used. In this design to minimize the pin count a digitally programmable delay element is designed and the control word is applied by the input interface through a serial input pin.
15 digitally programmable delay elements are cascaded and controlled by a thermometer code. Each digitally programmable delay element gives a delay of 2.2ps, resulting in a delay of 33ps over full thermometer control change for the LSBs.

For an n-bit counter implemented at the end of the start signal generation line, the maximum width of the calibration pulse is given by Eq. 7.1:

$$T_{\text{cal, max}} = 2^n \cdot T_{\text{start, min}} \,. \tag{7.1}$$

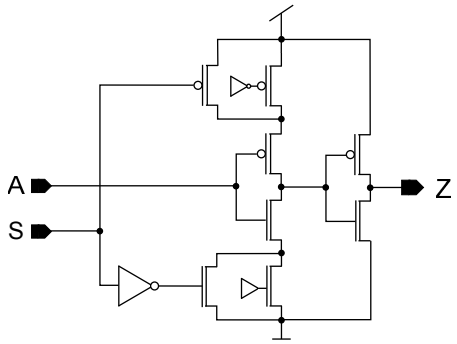


Figure 7.14.: Tunable delay element used in the TDC calibration circuit

Although the time interval between the start and the stop signal can be tuned in a coarse (MSBs) and a fine (LSBs) way, the absolute time resolution cannot be extracted from this information. Up to now the time measurements can determine the relative TDC behavior [99]. However, the absolute time interval corresponding to the output word of the TDC has to be measured separately [98].
For this purpose, the counters at the end of both the start and the stop signal generation lines are implemented (see Fig. 7.13). When the calibration-enable is activated, the start and stop circuitry form two ring oscillators. At the end of each delay line, a counter is inserted in order to measure the oscillations of the start and the stop signal. Outputs of the counters are used to calculate the time between the start and stop signals. During the known enabling time interval T_{cal} of the calibration pulse, the counters are enabled. Since both the start and stop signal lines are implemented in ring-oscillator-like structures and contain each an odd number of inverting stages, the signals will oscillate and the counters will count the number of cycles of the respective full loop oscillations.
The ring-oscillator configuration is beneficial for insensitivity against high-to-low and low-to-high asymmetries [99]. It is preferable to have the oscillations running over a relatively long time in order to suppress any possible noise induced delay variations and to have a stable reference frequency [99].
Finally, the skew ΔT between the signals propagated through the start and the stop line is then given by [98, 99]

$$\Delta T_{start,\,stop} = \frac{T_{\text{cal}}}{2} \cdot \left[\left(\frac{1}{N_{\text{osc, stop}}}\right) - \left(\frac{1}{N_{\text{osc, start}}}\right)\right] \tag{7.2}$$

where $N_{\text{osc, start}}$ and $N_{\text{osc, stop}}$ are the number of oscillations in the start and stop ring oscillator structures when the calibration pulse is enabled, respectively. The absolute time difference $\Delta T_{start,\,stop}$ between the start and the stop signal corresponds to the TDC output word. The start and stop counter outputs are written on the output register bank and accessible during the read out phase by the output interface.

7.3. On-chip Control for Stress and Measurement Cycles

By using the stimuli applied to the test chip, four modes of operations are obtained: DC stress, AC stress, measurement and TDC calibration mode. During the measurement and the AC stress modes, the paths under test are stimulated. However, only in the measurement mode the timing slack monitors are enabled by the Monitor-Enable signal. During the DC stress mode, the paths experience a steady state DC stress.
Figure 7.15 shows three phases for each measurement. In the first step (setup

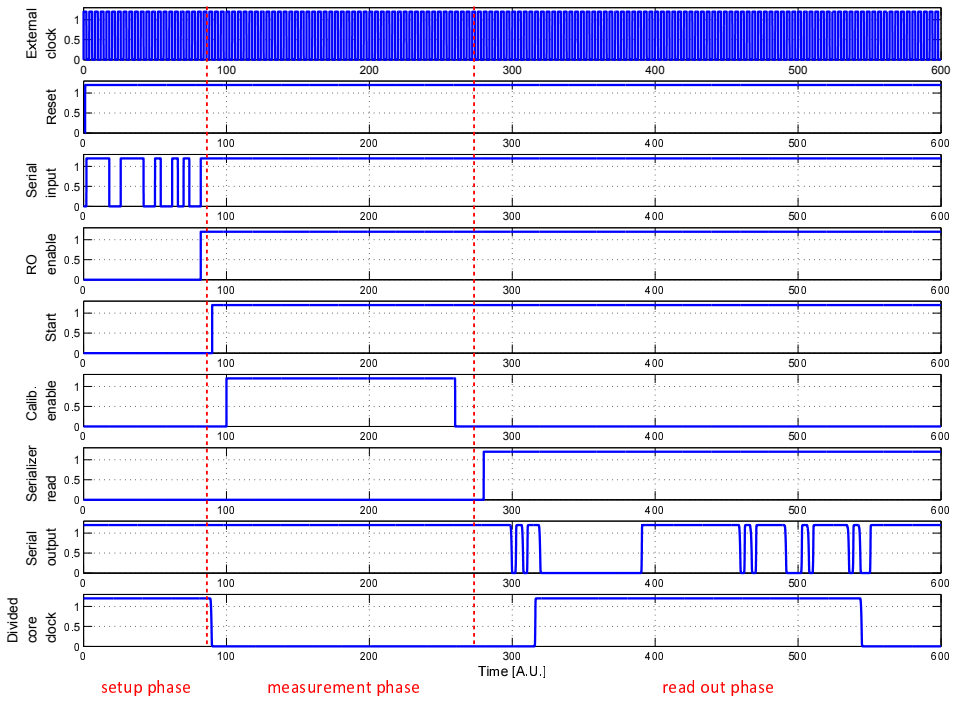


Figure 7.15.: Timing diagram in the measurement mode, all phases: setup, one-shot measurement, readout phase

phase) the control signals are uploaded to the test chip by the serial input, the external clock and the ring-oscillator-enable. The number of the path to be monitored (path-select), the control word for tuning the clock frequency as well as

the TDC calibration settings are uploaded in this phase. In the next phase, the one-shot measurement phase is initiated by the external start signal. If the start signal transitions to "1", all paths under test are enabled, i.e. oscillate triggered by the core clock signal. Afterward, the Monitor-Enable corresponding to the path-select will be enabled and the remaining timing slack of the path under test is captured by the in situ monitor. For the first CUT (see section 7.2), the TDC performs a one-shot measurement of the remaining timing slack of the path under test. Figure 7.16 shows the corresponding internal signals. For the second CUT

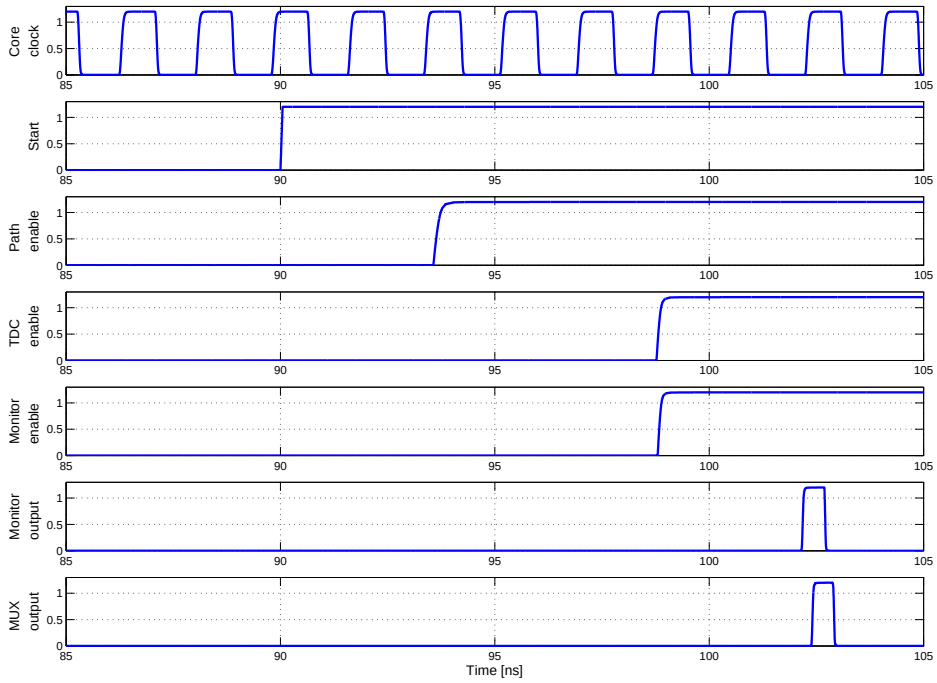


Figure 7.16.: Timing diagram in the measurement phase, in situ monitoring of the CUT by precise slack monitors

the remaining timing slack is converted to a digital code by the 2-bit in situ TDC monitors.
If the path-select is "0000" and a calibration mode is planned, the calibration-enable signal goes to high during the measurement phase to characterize the TDC. During the read out phase, the output-interface-control (serializer-read) is activated and the resulting measurement data are read out serially. The data to be read out includes the TDC measurement result, the output of the 2-bit in situ TDC monitors and the result of the calibration circuitry.
The following discusses all modes of operation for the test chips.

DC stress

In the setup phase the path-select is set to "0000" and the monitors are disabled. The start signal is set to "0" and does not transition to "1". Thus the paths are not oscillating as the path-enable signals connected to them remain zero. The circuit remains in the idle state in which the devices remain in a DC stress.

AC stress

The path-select is set to "0000" and the monitors are disabled. The start signal is set to "0" and transitions to "1". The circuit remains in a state in which all the paths are enabled and the circuitry experiences an AC stress.

Measurement, calibration of the TDC

The path-select is set to "0000" and the monitors are disabled. The start signal is set to "0" and transitions to "1". The circuit remains in a state waiting for the calibration-enable pulse. The occurrence of the calibration-enable pulse results in a single shot TDC measurement. After the calibration-enable changes to logic "0" the output interface provides the results to the output. Figure 7.17 shows the internal signals corresponding to this mode.

Measurement mode, in situ monitoring of the CUT

The path-select is set to the path number to be monitored. The start signal is set to "0" and then changes to "1". All the paths are enabled by the path-enable signal. The path under test is identified by the path-select in the input interface module. The Monitor-Enable of the monitor of the path to be monitored and the TDC-enable are activated after 3 clock cycles, as shown in Fig 7.16. Therefore, after a DC stress, the worst case degradation is monitored. A single shot measurement by the TDC is performed. Afterward, the output interface provides the results of the measurement to the output pad.

7.4. Overhead of the Monitoring System

Tables 7.2 and 7.3 show the evaluated overheads for the centralized and decentralized monitoring approaches , respectively. The values are evaluated based on the implemented paths under test in the chip. An exemplary circuit equipped with monitors at the end of 5% of the critical paths is assumed and the overheads for a such circuit are recalculated and shown in Tables 7.2 and 7.3. It should be noted that considering also the non-critical paths, the overheads will be even

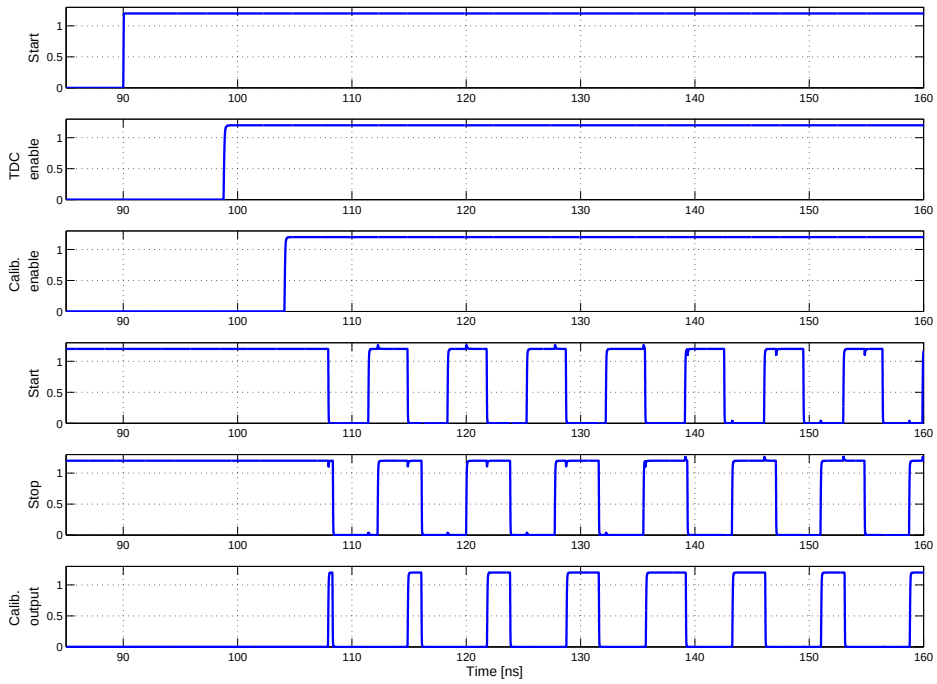


Figure 7.17.: Timing diagram in the measurement mode, calibration of the TDC

smaller. Moreover, the power overheads are evaluated for a certain clock frequency (500MHz) and assuming a 100% activity within the circuit under test. The 100% activity means that the monitors and the TDC are continuously switching. However, for an online monitoring system in a normal circuit with much less than 100% activity the power overheads drops dramatically as the switching activity of the paths would be much smaller than 100%. When the monitoring system is disabled, the power overhead drops below 1%.

7.5. Summary

In safety critical applications, in situ monitoring of the performance can be used for reliability assessment of the circuit under test. In this chapter the developed monitoring circuitry for a 40nm technology was discussed. The developed monitoring approach observes the effects of degradation mechanisms over the lifetime of the circuits. Thus, the developed approach provides a useful means to deliver highly reliable systems with extremely low error rates. By exploiting the result of the monitoring approach, near future failures can be predicted and maintenance

	area overhead	power overhead
monitor type 1 (Fig. 5.8)	0.3%	1.2%
monitor type 2 (Fig. 5.11)	0.7%	1.5%
monitor type 3 (Fig. 5.15)	0.4%	1.6%
MUX tree	2.0%	0.5%
TDC	8.3%	32.7%
Total with monitor type 1	8.7%	34.2%
Total with monitor type 2	9.1%	34.5%
Total with monitor type 3	8.8%	34.6%

Table 7.2.: Overheads of the centralized monitoring system with precise slack monitors for an exemplary circuit equipped with monitors at the end of 5% of the critical paths. The power overhead is evaluated when assuming a 100% activity rate for the paths under test and activated monitors and TDC.

	area overhead	power overhead
2-bit in situ TDC monitor	0.6%	2.8%
2-bit in situ TDC monitor (with additional delays)	0.7%	3.5%

Table 7.3.: Overheads of the decentralized monitoring system with 2-bit in situ TDC monitors for an exemplary circuit equipped with monitors at the end of 5% of the critical paths. The power overhead is evaluated when assuming a 100% activity rate for the paths under test and activated monitors. The decentralized monitoring system offers lower resolution with less overheads compared to the centralized approach.

before failure can be performed. Moreover, the extracted data can be used to adapt the operating parameters and perform a self-healing process for the device under test.

8. Summary and Outlook

In advanced technology nodes due to the strong silicon and system complexities, product reliability is at high risk. Especially, for safety critical applications ensuring high reliability requirements increases the costs (design, area, power etc.) tremendously. Thus, for such applications precise characterization of circuits during the design phase to predict the lifetime reliability is more demanding. Moreover, during the functional operation, predicting near-future failures is highly advantageous as it enables a reliability management during the entire lifetime of the circuit. Reliability assessment during both the design phase and the lifetime of the circuit reduces the costs and ensures that the application specific reliability requirements are satisfied. The scope of this thesis is the assessment of aging induced performance degradation during the design phase as well as through the lifetime of digital CMOS circuits.

The effect of different aging mechanisms in advanced CMOS technologies is discussed in chapter 2. Thereby, the effect of an aging induced threshold voltage shift on digital components is discussed. For the reliability assessment during the design phase, NBTI is regarded in this work as a dominant aging mechanism in 65nm and 40nm digital CMOS circuits. This thesis proposes to predict the timing degradation due to NBTI induced aging with a novel sufficiently accurate circuit level model, which considers both NBTI stress and recovery. The developed aging analysis tool is capable of analyzing complex synthesized as well as custom designed digital circuits. The aging analysis tool combines gate level and transistor level approaches. Thus, it is able to accurately characterize digital circuits with low computational effort compared to the transistor level approaches. To enable efficient integration of the models updated for new technologies, the physical model is separated from the circuit analysis in the aging analysis tool-set. A detailed description of the developed NBTI aging analysis tool-set and its applications to several test circuits is provided in chapter 3. As an example, the developed aging analysis tool is used to determine the reliability of a circuitry which generates a secure hash algorithm. Here, weak spots of the circuit as well as critical operating scenarios are identified.

The second part of this work focuses on monitoring the reliability status of the circuit during its lifetime. Thereby, timing properties of the circuit are observed by integrated in situ monitors as an indicator for the degradation level. These monitors provide timing information influenced by the current reliability status of the circuit under test. The monitoring methods and the system level considerations

are regarded in chapter 4. In this chapter online and offline monitoring concepts are compared and in both cases efficient placement of the in situ monitors within the circuit are discussed. In this context, a novel algorithm for the placement of the monitors in the online monitoring case is introduced.
In chapter 5 the developed in situ monitoring circuitry is discussed and evaluated through extensive simulations. In this chapter developed designs of different in situ monitors are discussed. Monitors observe the remaining timing slack of the path under test. Different approaches with different timing resolution for measuring the remaining slack of the paths equipped with monitors are discussed. Two main approaches for conversion of the extracted timing information to the digital domain are categorized as centralized and decentralized. In a decentralized approach the output of the monitors is a binary value. As an example, the developed low power/area 2 bit in situ monitors convert the remaining timing slack to a binary value with 3 timing thresholds. A centralized approach uses precise slack monitors which extract the remaining timing slack as a pulse. Afterward, a time to digital converter converts the extracted pulse to a digital value.
Two main safety critical applications to be equipped with the monitoring system are regarded in this thesis. Chapters 6 and chapter 7 demonstrate the development of the application specific monitoring systems as well as the results, respectively. As the first application, the monitoring system is implemented and evaluated in a chip to be used as a medical implant. Thus, the digital front-end of a neural measurement system implanted within patient's brain is equipped with the situ monitoring system. The efficiency and accuracy of the monitoring system is evaluated by simulations and experiments. In chapter 6 in situ monitoring is used for the reliability diagnosis in combination with adaptive voltage scaling for the neural measurement system. The monitoring system enables dynamic adaptation of operating parameters such as supply voltage and thus decreases the power consumption of the chips. Thus, the operating temperature of the implants is controlled and the potential health risks to the patients are reduced. Moreover, by adaptive voltage scaling, device aging is reduced and the lifetime of the circuits is prolonged by adapting the supply voltage to a slightly higher value in case of degradation. This ensures that the timing specifications are fulfilled. Even by assuming a pessimistic aging of 5%, applying the adaptive voltage scaling approach gives a power saving of 37%.
As the second application, a digital ASIC is chosen. In chapter 7 in situ monitors are implemented for characterizing the circuit under test. Both the above mentioned centralized approach and the decentralized approach are implemented in a demonstrator. For applications with high reliability requirements the in situ monitoring system provides a useful means for maintenance before failure as well as for a self-healing process through taking countermeasures such as frequency scaling. As an example the area and power overheads for a circuit equipped with 2 bit in situ monitors with an exemplary percentage of 5% of the critical paths

is estimated as 0.6% and 2.8%, respectively. The experimental results of the test chip will be presented within a follow-up project.
The implemented monitoring system offers low power and low area overheads in combination with high accuracy. Moreover, the monitoring system enables application of adaptive voltage and/or frequency scaling methods which in turn reduce the aging and increase the lifetime of the circuits equipped with the monitors. The extracted information regarding the reliability status of the circuit under test can also be used to predict near future failures. Thus, devices can be either substituted (maintenance before failure) and/or the mentioned adaptive methods can be carried out. The quantitative evaluations by simulation and experimental data prove the applicability and the benefits of the monitoring concept for highly reliable systems.
Nowadays, where electronic devices find their way into every aspect of our lives, might it be in smart home or automatic driving, demands for the highly reliable systems is further increasing. Thus, observability of the circuit reliability by an in situ monitoring system enables to take countermeasures (when necessary) and thus can have considerable contributions to the safety of the human lives.

A. List of Symbols

$\boldsymbol{C}$	correlation matrix for the capture and emission mechanisms in the energy domain
ΔI_d	induced shift of the transistor drain current
ΔV_{th}	induced shift of the transistor threshold voltage
$\Delta V_{th,R}$	induced shift of the transistor threshold voltage by the removable NBTI
$\Delta V_{th,P}$	induced shift of the transistor threshold voltage by the permanent NBTI
E_{Ac}	capture activation energy
E_{Ae}	emission activation energy
E_F	Fermi energy
E_{GL}	set of interconnect edges
E_V	energy of the valence band
FF	output of the flip-flop
G_{TL}	transistor level graph
I_d	drain current
I_{D0}	drain current at $\lvert V_{GS}\rvert = V_{DD}$
k_B	Boltzmann's constant
$k_{i,j}$	transition probability per unit time
λ	scaling factor
$\mu_{c,e}$	vector of mean activation energies for capture and emission mechanisms
$NBTI_P$	permanent NBTI

$NBTI_{P,R}$	permanent and recoverable NBTI
$NBTI_R$	recoverable NBTI
n_{pre}	count of pre-errors
$n_{stress-relax}$	number of stress relaxation cycles
N_V	effective density of states in the valence band
p	density of holes
Φ_R	regular bivariate normal distribution describing the recoverable NBTI
Φ_P	regular bivariate normal distribution describing the permanent NBTI
$\rho_{c,e}$	correlation factor between the activation energies for charge capture and emission
σ_{cs}	capture/emission cross section
σ_c	standard deviation of the activation energies for charge capture mechanism
σ_e	standard deviation of the activation energies for charge emission mechanism
SiO_2	silicon dioxide
t_{AC}	overall stress time
τ_c	capture time constant
τ_e	emission time constant
T_{AC}	AC stress period
T_{clk}	clock period
T_{cq}	clock-to-Q delay of the flip-flop
$t_{d,\,max}$	maximum delay within the circuit
$t_{d,\,crt}$	critical transition time for a path
$Temp$	temperature

T_{jitter}	clock jitter
t_M	measurement delay
t_{relax}	relaxation time
t_{stress}	stress time
T_{pd}	propagation delay through the combinatorial path
T_{setup}	setup time of the flip-flop
T_{skew}	clock skew
t_T	transition time
V_{DD}	supply voltage
V_{DS}	drain source voltage of the transistor
V_{GS}	gate source voltage of the transistor
V_{stress}	stress voltage
V_{TL}	set of transistor vertices

B. List of Abbreviations

ATPG	automatic test pattern generator
AVS	adaptive voltage scaling
BIST	built in self test
BTI	bias temperature instability
CET	capture emission time
CHCI	conductive hot carrier injection
CMOS	complementary metal oxide semiconductor
CUT	circuit under test
DFT	design for testability
DNL	differential non-linearity
DUF	duty factor
DUT	device under test
ECoG	electrocorticography
ECU	electronic control unit
EES	electron-electron scattering
FPGA	field-programmable gate array
GRO	gated ring oscillator
GSM	global system for mobile communications
HCI	hot carrier injection
HDL	hardware description language
HTOL	high temperature operation life test

IC	integrated circuit
LoC	launch-on-capture
LoS	launch-on-shift
LSB	least significant bit
LUT	look-up table
ML	master latch output
MO	monitor output
MOSFET	metal-oxide-semiconductor field effect transistor
MSB	most significant bit
MVE	multiple vibrational excitation
MUX	multiplexer
NBTI	negative bias temperature instability
NCHCI	non conducting hot carrier injection
NMOS	negative channel metal-oxide-semiconductor field effect transistor
NMP	nonradiative multiphonon process
NMS	neural measurement system
PBC	pseudo-thermometer to binary converter
PBTI	positive bias temperature instability
PLL	phase locked loop
PMOS	positive channel metal-oxide-semiconductor field effect transistor
PP	potential path
PVTA	process, voltage, temperature and aging
RF	radio frequency
RO	ring oscillator
RTL	register-transfer level

SHA	secure hash algorithm
STA	static timing analysis
SA	switching activity
TDDB	time-dependent dielectric breakdown
TDC	time to digital converter
TDF	transition delay fault
TRC	tunable replica circuit
VHDL	VHSIC (very high speed integrated circuit) hardware description language

Bibliography

[1] B. Vaidyanathan and A. Oates, "Technology scaling effect on the relative impact of NBTI and process variation on the reliability of digital circuits," *IEEE Transactions on Device and Materials Reliability*, vol. 12, no. 2, pp. 428–436, June 2012.

[2] E. Maricau and G. Gielen, "Efficient reliability simulation of analog ICs including variability and time-varying stress," in *Design, Automation Test in Europe Conference Exhibition (DATE)*, 2009, pp. 1238–1241.

[3] S. Kupke, S. Knebel, S. Rahman, S. Slesazeck, T. Mikolajick, R. Agaiby, and M. Trentzsch, "Dynamic off-state TDDB of ultra short channel HKMG nFETS and its implications on CMOS logic reliability," in *IEEE International Reliability Physics Symposium*, June 2014, pp. 5B.1.1–5B.1.6.

[4] E. Mintarno, J. Skaf, R. Zheng, J. Velamala, Y. Cao, S. Boyd, R. Dutton, and S. Mitra, "Optimized self-tuning for circuit aging," in *Design, Automation Test in Europe Conference Exhibition (DATE), 2010*, March 2010, pp. 586–591.

[5] N. Pour Aryan, A. Listl, L. Heiss, C. Yilmaz, G. Georgakos, and D. Schmitt-Landsiedel, "From an analytic NBTI device model to reliability assessment of complex digital circuits," in *IEEE 20th International On-Line Testing Symposium (IOLTS)*, July 2014, pp. 19–24.

[6] N. Pour Aryan, G. Georgakos, and D. Schmitt-Landsiedel, "Reliability monitoring of digital circuits by in situ timing measurement," in *23rd International Workshop on Power and Timing Modeling, Optimization and Simulation (PATMOS)*, Sept 2013, pp. 150–156.

[7] N. Pour Aryan, N. Heidmann, M. Wirnshofer, N. Hellwege, J. Pistor, D. Peters-Drolshagen, G. Georgakos, S. Paul, and D. Schmitt-Landsiedel, "Power efficient digital IC design for a medical application with high reliability requirements," in *24th International Workshop on Power and Timing Modeling, Optimization and Simulation (PATMOS)*, 2014.

[8] M. Wirnshofer, *Variation-aware adaptive voltage scaling for digital CMOS circuits*. Springer Series in Advanced Microelectronics, 2013.

[9] M. Wirnshofer, L. Heiss, A. Kakade, N. Aryan, G. Georgakos, and D. Schmitt-Landsiedel, "Adaptive voltage scaling by in-situ delay monitoring for an image processing circuit," in *IEEE 15th International Symposium on Design and Diagnostics of Electronic Circuits and Systems (DDECS)*, 2012, pp. 205–208.

[10] F. Ahmed and L. Milor, "Reliable cache design with on-chip monitoring of NBTI degradation in SRAM cells using BIST," in *28th VLSI Test Symposium (VTS)*, April 2010, pp. 63–68.

[11] S. Khan and S. Hamdioui, "Modeling and mitigating NBTI in nanoscale circuits," in *IEEE 17th International On-Line Testing Symposium (IOLTS)*, July 2011, pp. 1–6.

[12] H. Luo, Y. Wang, J. Velamala, Y. Cao, Y. Xie, and H. Yang, "The impact of correlation between NBTI and TDDB on the performance of digital circuits," in *IEEE 54th International Midwest Symposium on Circuits and Systems (MWSCAS)*, aug. 2011, pp. 1 –4.

[13] D. Lorenz, G. Georgakos, and U. Schlichtmann, "Aging analysis of circuit timing considering NBTI and HCI," in *15th IEEE International On-Line Testing Symposium (IOLTS)*, june 2009, pp. 3 –8.

[14] T. Grasser, B. Kaczer, H. Reisinger, P.-J. Wagner, and M. Toledano-Luque, "Recent developments in understanding the bias temperature instability," in *28th International Microelectronics Conference (MIEL)*, pp. 315–322, 2012.

[15] H. Reisinger, T. Grasser, K. Ermisch, H. Nielen, C. Gustin, and C. Schluender, "Understanding and modeling AC BTI," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. 597-604, 2011.

[16] C. Ma, H. Mattausch, M. Miyake, T. Iizuka, M. Miura-Mattausch, K. Matsuzawa, S. Yamaguchi, T. Hoshida, M. Imade, R. Koh, T. Arakawa, and J. He, "Compact reliability model for degradation of advanced p-MOSFETs due to NBTI and hot-carrier effects in the circuit simulation," in *IEEE International Reliability Physics Symposium (IRPS)*, April 2013, pp. 2A.3.1–2A.3.6.

[17] V. Huard, "Two independent components modeling for negative bias temperature instability," in *IEEE International Reliability Physics Symposium (IRPS)*, May 2010, pp. 33–42.

[18] H. Oner, B. Bayrakci, and Y. Leblebici, "A compact monitoring circuit for real-time on-chip diagnosis of hot-carrier induced degradation," in *IEEE*

International Conference on Microelectronic Test Structures (ICMTS), 1997, pp. 72 –76.

[19] J. Tschanz, K. Bowman, S. Walstra, M. Agostinelli, T. Karnik, and V. De, "Tunable replica circuits and adaptive voltage-frequency techniques for dynamic voltage, temperature, and aging variation tolerance," in *Symposium on VLSI Circuits*, 2009, pp. 112 –113.

[20] T. Iizuka, T. Nakura, and K. Asada, "Buffer-ring-based all-digital on-chip monitor for PMOS and NMOS process variability and aging effects," in *IEEE 13th International Symposium on Design and Diagnostics of Electronic Circuits and Systems (DDECS)*, 2010, pp. 167 –172.

[21] T.-H. Kim, R. Persaud, and C. Kim, "Silicon odometer: An on-chip reliability monitor for measuring frequency degradation of digital circuits," in *IEEE Symposium on VLSI Circuits*, june 2007, pp. 122 –123.

[22] M. Eireiner, S. Henzler, G. Georgakos, J. Berthold, and D. Schmitt-Landsiedel, "In-situ delay characterization and local supply voltage adjustment for compensation of local parametric variations," *IEEE Journal of Solid-State Circuits*, vol. 42, no. 7, pp. 1583–1592, 2007.

[23] M. Agarwal, V. Balakrishnan, A. Bhuyan, K. Kim, B. C. Paul, W. Wang, B. Yang, Y. Cao, and S. Mitra, "Optimized circuit failure prediction for aging: Practicality and promise," in *Proc. IEEE Int. Test Conf. ITC 2008*, 2008, pp. 1–10.

[24] A. Amouri and M. Tahoori, "A low-cost sensor for aging and late transitions detection in modern FPGAs," in *International Conference on Field Programmable Logic and Applications (FPL)*, sept. 2011, pp. 329 –335.

[25] *The International Technology Roadmap for Semiconductors (ITRS): Design*, [Online] Available at: http://www.itrs.net/Links/2011ITRS/2011Chapters/2011Design.pdf, 2011.

[26] A. Strong, E. Wu, R. Vollertsen, J. Sune, G. Rosa, T. Sullivan, and S. Rauch, *Reliability Wearout Mechanisms in Advanced CMOS Technologies*, ser. IEEE Press Series on Microelectronic Systems. Wiley, 2009.

[27] T. Grasser, "Stochastic charge trapping in oxides: From random telegraph noise to bias temperature instabilities," in *International Reliability Physics Symposium (IRPS)*, pp. 39-70, 2011.

[28] D. Ioannou, S. Mittl, and G. La Rosa, "Positive bias temperature instability effects in nMOSFETs with HfO2/TiN gate stacks," *IEEE Transactions on Device and Materials Reliability*, vol. 9, no. 2, pp. 128–134, June 2009.

[29] F. Ahmed and L. Milor, "Ring oscillator based embedded structure for decoupling PMOS/NMOS degradation with switching activity replication," in *IEEE International Conference on Microelectronic Test Structures (ICMTS)*, 2010.

[30] E. Maricau and G. Gielen, *Analog IC Reliability in Nanometer CMOS*, ser. Analog Circuits and Signal Processing. Springer, 2013.

[31] C. Hu, S. C.Tam, F.-C. Hsu, P.-K. Ko, T.-Y. Chan, and K. Terrill, "Hot-electron-induced MOSFET degradation - model, monitor, and improvement," *IEEE Journal of Solid-State Circuits*, vol. 20, no. 1, pp. 295–305, Feb 1985.

[32] C. Hu, "Lucky-electron model of channel hot electron emission," in *International Electron Devices Meeting*, vol. 25, 1979, pp. 22–25.

[33] S. Rauch, G. La Rosa, and F. J. Guarin, "Role of E-E scattering in the enhancement of channel hot carrier degradation of deep-submicron NMOSFETs at high VGS conditions," *IEEE Transactions on Device and Materials Reliability*, vol. 1, no. 2, pp. 113–119, Jun 2001.

[34] K. Hess, B. Tuttle, F. Register, and D. Ferry, "Magnitude of the threshold energy for hot electron damage in metal oxide semiconductor field effect transistors by hydrogen desorption," *Applied Physics Letters*, vol. 75, pp. 3147–3149, 1999.

[35] D. Lorenz, M. Barke, and U. Schlichtmann, "Aging analysis at gate and macro cell level," in *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2010.

[36] E. Amat, R. Rodriguez, M. Nafria, X. Aymerich, T. Kauerauf, R. Degraeve, and G. Groeseneken, "New insights into the wide ID range channel hot-carrier degradation in high-k based devices," in *IEEE International Reliability Physics Symposium*, April 2009, pp. 1028–1032.

[37] S. Sahhaf, R. Degraeve, P. Roussel, T. Kauerauf, B. Kaczer, and G. Groeseneken, "TDDB reliability prediction based on the statistical analysis of hard breakdown including multiple soft breakdown and wear-out," in *IEEE International Electron Devices Meeting (IEDM)*, Dec 2007, pp. 501–504.

[38] N. Suzumura, M. Ogasawara, T. Furuhashi, and T. Koyama, "Study on vertical TDDB degradation mechanism and its relation to lateral TDDB in Cu/low-k damascene structures," in *IEEE International Reliability Physics Symposium*, June 2014, pp. 3A.4.1–3A.4.6.

[39] K. Martin, *Digital Integrated Circuit Design*, ser. Oxford series in electrical and computer engineering. Oxford University Press, 2000.

[40] M. White, *A Study of Nanometer Semiconductor Scaling Effects on Microelectronics Reliability.* University of Maryland, College Park. Mechanical Engineering, 2009.

[41] 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*, vol. 25, no. 2, pp. 584–594, Apr. 1990.

[42] S. Kumar, C. Kim, and S. Sapatnekar, "NBTI-aware synthesis of digital circuits," in *44th ACM/IEEE Design Automation Conference*, pp. 370–375, 2007.

[43] K. Saluja, S. Vijayakumar, W. Sootkaneung, and X. Yang, "NBTI degradation: A problem or a scare," in *21st International Conference on VLSI Design (VLSID)*, pp. 137–142, 2008.

[44] P. Bipul, K. Kang, H. Kufluoglu, M. Alam, and K. Roy, "Impact of NBTI on the temporal performance degradation of digital circuits," in *IEEE Electron Device Letters*, pp. 560–562, 2005.

[45] V. Huard, C. Parthasarathy, A. Bravaix, C. Guerin, and E. Pion, "CMOS device design-in reliability approach in advanced nodes," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. 624–633, 2009.

[46] B. Peng, I.-Y. Chen, S.-Y. Kuo, and C. Bolger, "IC HTOL test stress condition optimization," in *19th IEEE International Symposium on Defect and Fault Tolerance (DFT) in VLSI Systems*, Oct 2004, pp. 272–279.

[47] M. Cho, N. Sathe, A. Raychowdhury, and S. Mukhopadhyay, "Optimization of burn-in test for many-core processors through adaptive spatiotemporal power migration," in *Test Conference (ITC), 2010 IEEE International*, Nov 2010, pp. 1–9.

[48] R. Kwasnick, M. Reilly, J. Hatfield, S. Johnson, and A. Rahman, "Impact of VLSI technology scaling on HTOL," in *Reliability Physics Symposium (IRPS), 2012 IEEE International*, April 2012, pp. 5C.3.1–5C.3.5.

[49] D. Crowe and A. Feinberg, *Design for Reliability*, ser. Electronics Handbook Series. Taylor & Francis, 2001.

[50] N. Pour Aryan, M. Wirnshofer, S. Aghaie, G. Georgakos, and D. Schmitt-Landsiedel, "Timing slack monitoring for reliability diagnosis," in *edaWorkshop13, Electronic Design Automation (EDA)*, 2013.

[51] C. Yilmaz, L. Heiss, C. Werner, and D. Schmitt-Landsiedel, "Modeling of NBTI-recovery effects in analog CMOS circuits," in *Reliability Physics Symposium (IRPS), 2013 IEEE International*, 2013.

[52] A. Islam, H. Kufluoglu, D. Varghese, S. Mahapatra, and M. Alam, "Recent issues in negative bias temperature instability: Initial degradation, field dependence of interface trap generation, hole trapping effects, and relaxation," *IEEE Transactions on Electron Devices*, vol. 54, no. 9, pp. 2143,2154, Sept 2007.

[53] N. Gong, R. Wang, C. Liu, J. Zou, and R. Huang, "On the AC random telegraph noise (RTN) in MOS devices: An improved multi-phonon based model," in *IEEE 11th International Conference on Solid-State and Integrated Circuit Technology (ICSICT)*, Oct 2012, pp. 1–3.

[54] F. Schanovsky, O. Baumgartner, V. Sverdlov, and T. Grasser, "A multi scale modeling approach to non-radiative multi phonon transitions at oxide defects in MOS structures," *Journal of Computational Electronics*, vol. 11, pp. 218–224, 2012.

[55] S. Pantelides, *The Physics of SiO2 and Its Interfaces: Proceedings of the International Topical Conference on the Physics of SiO2 and Its Interfaces*. Elsevier Science, 2013. [Online]. Available: https://books.google.de/books?id=WzQvBQAAQBAJ

[56] T. Grasser, P.-J. Wagner, H. Reisinger, T. Aichinger, G. Pobegen, M. Nelhiebel, and B. Kaczer, "Analytic modeling of the bias temperature instability using capture/emission time maps," in *IEEE International Electron Devices Meeting (IEDM)*, pp. 618–621, 2011.

[57] T. Grasser, T. Aichinger, G. Pobegen, H. Reisinger, P. Wagner, J. Franco, M. Nelhiebel, and B. Kaczer, "The 'permanent' component of NBTI: Composition and annealing," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. 605–613, 2011.

[58] T. Grasser, B. Kaczer, P. Hehenberger, W. Goes, R. O'Connor, H. Reisinger, C. Gustin, and C. Schluender, "Simultaneous extraction of recoverable and permanent components contributing to bias-temperature instability," in *International Electron Devices Meeting (IEDM)*, pp. 801–3804, 2007.

[59] N. Jha, P. Reddy, D. Sharma, and V. Ramgopal Rao, "NBTI degradation and its impact for analog circuit reliability," *IEEE Transactions on Electron Devices*, vol. 52, no. 12, pp. 2609–2615, Dec 2005.

[60] J. Martin-Martinez, M. Moras, N. Ayala, V. Velayudhan, R. Rodriguez, M. Nafria, and X. Aymerich, "Modeling of time-dependent variability caused by bias temperature instability," in *Spanish Conference on Electron Devices (CDE)*, Feb 2013, pp. 241–244.

[61] V. Huard, C. Parthasarathy, A. Bravaix, C. Guerin, and E. Pion, "CMOS device design-in reliability approach in advanced nodes," in *IEEE International Reliability Physics Symposium (IRPS)*, 2009, pp. pp. 624–633,.

[62] R. Tu, E. Rosenbaum, W. Chan, C. Li, E. Minami, K. Quader, and P. C. H. Ko, "Berkeley Reliability Tools - BERT," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, pp. 1524 - 1534 , 1993.

[63] "Reliability simulation in integrated circuit design," White Paper, Cadence Design Systems, Inc., Tech. Rep., 2003.

[64] C. Liu, P. Ren, R. Wang, R. Huang, J. Ou, Q. Huang, J. Zou, J. Wang, J. Wu, S. Yu, H. Wu, S.-W. Lee, and Y. Wang, "New observations on AC NBTI induced dynamic variability in scaled high-k Metal-gate MOSFETs: Characterization, origin of frequency dependence, and impacts on circuits," in *IEEE International Electron Devices Meeting (IEDM)*, Dec 2012, pp. 19.5.1–19.5.4.

[65] S. Desai, S. Mukhopadhyay, N. Goel, N. Nanaware, B. Jose, K. Joshi, and S. Mahapatra, "A comprehensive AC / DC NBTI model: Stress, recovery, frequency, duty cycle and process dependence," in *IEEE International Reliability Physics Symposium (IRPS)*, April 2013, pp. XT.2.1–XT.2.11.

[66] H. Reisinger, T. Grasser, W. Gustin, and C. Schlünder, "The statistical analysis of individual defects constituting NBTI and its implications for modeling DC- and AC-stress," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. 7–15, 2010.

[67] D. Dobberpuhl and R. Witek, "A 200MHz 64b diual-issue CMOS microprocessor," *IEEE International Solid-State Circuits Conference*, 1992.

[68] J. Kathuria, M. Ayoubkhan, and A. Noor, "A review of clock gating techniques," *MIT International Journal of Electronics and Communication Engineering*, 2011.

[69] A. Chakraborty and D. Pan, "Skew management of NBTI impacted gated clock trees," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2013.

[70] M. Wolf, A. Weimerskirch, and C. Paar, *Secure In-Vehicle Communication*. Springer, 2006.

[71] H. Casier, P. Moens, and K. Appeltans, "Technology considerations for automotive (automotive electronics)," in *30th European Solid-State Circuits Conference (ESSCIRC)*, 2004.

[72] K. Han, S. Potluri, and K. Shin, "On authentication in a connected vehicle: Secure integration of mobile devices with vehicular networks," in *International Conference on Cyber-Physical Systems (ICCPS)*, 2013, pp. 160–169.

[73] P. Leteinturier, "Automotive semi-conductor trend challenges," in *Design, Automation and Test in Europe (DATE)*, vol. 1, March 2006, pp. 1–1.

[74] F. Sagstetter, M. Lukasiewycz, S. Steinhorst, M. Wolf, A. Bouard, S. Harris, W. .and Jha, T. Peyrin, A. Poschmann, and S. Chakraborty, "Security challenges in automotive hardware/software architecture design," in *Design, Automation Test in Europe Conference Exhibition (DATE)*, 2013.

[75] K. Koscher, A. Czeskis, F. Roesner, S. Patel, T. Kohno, S. Checkoway, D. McCoy, B. Kantor, D. Anderson, H. Shacham, and S. Savage, "Experimental security analysis of a modern automobile," in *IEEE Symposium on Security and Privacy (SP)*, 2010, pp. 447 – 462,.

[76] Z. Jian and J. Xuling, "Encryption system design based on DES and SHA-1," in *11th International Symposium on Distributed Computing and Applications to Business, Engineering Science (DCABES)*, 2012.

[77] N. Devtaprasanna, A. Gunda, P. Krishnamurthy, S. Reddy, and I. Pomeranz, "Test generation for open defects in CMOS circuits," in *21st IEEE International Symposium on Defect and Fault Tolerance (DFT) in VLSI Systems*, Oct 2006, pp. 41–49.

[78] S. Wang, "Generation of low power dissipation and high fault coverage patterns for scan-based BIST," in *International Test Conference*, 2002, pp. 834–843.

[79] D. Xiang, M.-J. Chen, J.-G. Sun, and H. Fujiwara, "Improving test effectiveness of scan-based BIST by scan chain partitioning," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, no. 6, pp. 916–927, June 2005.

[80] I. Pomeranz and S. Reddy, "Scan-based delay fault tests for diagnosis of transition faults," in *21st IEEE International Symposium on Defect and Fault Tolerance in VLSI Systems*, Oct 2006, pp. 419–427.

[81] M. Geilert, J. Alt, and M. Zimmermann, "On the efficiency of the transition fault model for delay faults," in *IEEE International Conference on Computer-Aided Design (ICCAD), Digest of Technical Papers*, Nov 1990, pp. 272–275.

[82] S. Cremoux, C. Fagot, P. Girard, C. Landrault, and S. Pravossoudovitch, "A new test pattern generation method for delay fault testing," in *14th VLSI Test Symposium*, Apr 1996, pp. 296–301.

[83] A. Singh and G. Xu, "Output hazard-free transition tests for silicon calibrated scan based delay testing," in *24th IEEE VLSI Test Symposium*, April 2006, pp. 7 pp.–357.

[84] J. Savir and S. Patil, "On broad-side delay test," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 2, no. 3, pp. 368–372, Sept 1994.

[85] I. Park and E. McCluskey, "Launch-on-shift-capture transition tests," in *Test Conference, 2008. ITC 2008. IEEE International*, Oct 2008, pp. 1–9.

[86] E. Koser, N. Pour Aryan, M. Wirnshofer, G. Georgakos, W. Stechele, and D. Schmitt-Landsiedel, "RELY - Reliability of SoCs for safety critical applications," in *edaWorkshop 12, Electronic Design Automation (EDA)*, 2012.

[87] N. Pour Aryan, M. Wirnshofer, G. Georgakos, and D. Schmitt-Landsiedel, "An in situ timing measurement method for reliability diagnosis of digital circuits," in *ITG/GI/GMM-Fachtagung Zuverlaessigkeit und Entwurf (ZuE)*, 2013.

[88] M. Abramovici and C. Stroud, "BIST-based delay-fault testing in FPGAs," in *8th IEEE International On-Line Testing Workshop*, 2002, pp. 131–134.

[89] I. Harris, P. Menon, and R. Tessier, "BIST-based delay path testing in FPGA architectures," in *International Test Conference*, 2001, pp. 932–938.

[90] M. Wirnshofer, N. P. Aryan, L. Heiss, G. Georgakos, and D. Schmitt-Landsiedel, "On-Line supply voltage scaling based on in situ delay monitoring to adapt for PVTA variations," *Journal of Circuits, Systems, and Computers*, 2012.

[91] N. Pour Aryan, L. Heiß, D. Schmitt-Landsiedel, G. Georgakos, and M. Wirnshofer, "Comparison of in-situ delay monitors for use in adaptive voltage scaling," *Advances in Radio Science*, vol. 10, pp. 215–220, 2012.

[92] B. Kumar, N. Sharma, K. Kishore, and A. Rajakumari, "Variable input delay CMOS logic for dynamic IR drop reduction," in *Asia Pacific Conference on Postgraduate Research in Microelectronics and Electronics (PrimeAsia)*, 2012, pp. 79–84.

[93] S. Das, C. Tokunaga, S. Pant, W. H. Ma, S. Kalaiselvan, K. Lai, D. M. Bull, and D. T. Blaauw, "RazorII: In situ error detection and correction for PVT and SER tolerance," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 1, pp. 32–48, 2009.

[94] K. Shi and D. Howard, "Challenges in sleep transistor design and implementation in low-power designs," in *43rd ACM/IEEE Design Automation Conference*, 2006, pp. 113–116.

[95] M. Wirnshofer, L. Heiss, G. Georgakos, and D. Schmitt-Landsiedel, "An energy-efficient supply voltage scheme using in-situ Pre-Error detection for on-the-fly voltage adaptation to PVT variations," in *13th International Symposium on Integrated Circuits (ISIC)*, 2011, pp. 94–97.

[96] M. Maymandi-Nejad and M. Sachdev, "A digitally programmable delay element: design and analysis," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 11, no. 5, pp. 871–878, 2003.

[97] M. Straayer and M. Perrott, "An efficient high-resolution 11-bit noise-shaping multipath gated ring oscillator tdc," in *VLSI Circuits, 2008 IEEE Symposium on*, June 2008, pp. 82–83.

[98] S. Henzler, S. Koeppe, D. Lorenz, W. Kamp, R. Kuenemund, and D. Schmitt-Landsiedel, "A local passive time interpolation concept for variation-tolerant high-resolution time-to-digital conversion," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 7, pp. 1666–1676, July 2008.

[99] S. Henzler, *Time-to-Digital Converters*. Springer Series in Advanced Microelectronics 29, Springer, 2010.

[100] J. Pistor, J. Hoeffmann, D. Rotermund, E. Tolstosheeva, T. Schellenberg, D. Boll, V. Gordillo-Gonzalez, S. Mandon, D. Peters-Drolshagen, A. Kreiter, M. Schneider, W. Lang, K. Pawelzik, and S. Paul, "Development of a fully implantable recording system for ECoG signals," in *Design, Automation Test in Europe Conference Exhibition (DATE)*, March 2013, pp. 893–898.

[101] N. Heidmann, N. Hellwege, T. Hohlein, T. Westphal, D. Peters-Drolshagen, and S. Paul, "Modeling of an analog recording system design for ECoG and AP signals," in *Design, Automation and Test in Europe Conference and Exhibition (DATE)*, March 2014, pp. 1–6.

[102] J. Pistor, J. Hoeffmann, D. Peters-Drolshagen, and S. Paul, "A programmable neural measurement system for spikes and local field potentials," in *Symposium on Design, Test, Integration and Packaging of MEMS/MOEMS (DTIP)*, May 2011, pp. 200–205.

[103] K. Hirairi, Y. Okuma, H. Fuketa, T. Yasufuku, M. Takamiya, M. Nomura, H. Shinohara, and T. Sakurai, "1340nm CMOS by adaptive power supply voltage control with parity-based error prediction and detection (PEPD) and fully integrated digital LDO," in *IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, Feb 2012, pp. 486–488.

[104] A. Aal and T. Polte, "On component reliability and system reliability for automotive applications," in *IEEE International Integrated Reliability Workshop Final Report (IRW)*, Oct 2012, pp. 168–170.

[105] E. Ferrati, "The reliability of the integrated circuits in automotive industry," in *IEEE International Workshop on Defect and Fault Tolerance in VLSI Systems*, Oct 1993, pp. 125–126.

[106] P. Hupfer, "Complex electronics for automobiles - what is left for the driver to do?" *Tagung Elektronik im Kraftfahrzeug der VDI-Gesellschaft Fahrzeug- und Verkehrstechnik*, 2001.

[107] R. Rivett, "Hazard identification and classification: ISO26262- the application of IEC61505 to the automotive sector," in *5th IET Seminar on SIL Determination*, Dec 2009, pp. 1–24.

[108] ISO 26262-1:2011 Road vehicles – Functional safety. International Organization for Standardization. [Online]. Available: www.iso.org