Power-efficient application of Sleep Transistors to enhance the reliability of integrated circuits

Claas Cornelius[1]*, Frank Sill Torres[2], and Dirk Timmermann[1]

[1]Inst. of Applied Microelectronics and Computer Engineering, University of Rostock, Rostock, 18051, Germany, {claas.cornelius, dirk.timmermann}@uni-rostock.de

[2]Dept. of Electronic Engineering, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, 31270-010, Brazil, franksill@ufmg.br

* corresponding author: Claas Cornelius

Address:
University of Rostock
Institute of Applied Microelectronics and Computer Engineering
House 1, Room 1332
Richard-Wagner-Str. 31
18119 Rostock-Warnemünde, Germany

Office : (+49) 381 / 498-7278
Fax : (+49) 381 / 498-118 7251
Email : claas.cornelius@uni-rostock.de

Date of Receiving: **to be completed by the Editor**

Date of Acceptance: **to be completed by the Editor**

Power-efficient application of Sleep Transistors to enhance the reliability of integrated circuits

Claas Cornelius, Frank Sill Torres and Dirk Timmermann

Abstract *— CMOS is furthermore the most widespread technology for digital designs as no feasible alternative is in sight to date and in the near future. The fundamental causes for this supremacy so far are the capability for miniaturization as well as the reliability and robustness of CMOS. Against the background of nanotechnology though, reliability concerns are arising with an alarming pace. The consequence is an increasing demand for approaches to improve both yield and lifetime reliability of today's complex integrated systems. Hence, a common solution is the redundant implementation of components. However, redundancy contradicts those other efforts in order to cope with power dissipation. Thus, the essential contribution of this work is an approach that increases the lifetime reliability of integrated circuits while delay and power penalties are kept to a minimum. Accordingly, "Sleep Transistors" (as a common technique to reduce standby leakage) are combined with the idea of modular redundancy. Furthermore, we propose an extended flow for the quantification of reliability on transistor level. Finally, the presented simulation results evidence that the suggested approach increases the lifetime reliability by more than a factor of two compared to initial designs.*

Keywords — Reliability; Sleep Transistors; Redundancy; Simulation, Modeling

1 INTRODUCTION

The continuous scaling of technology has enabled the exceptional improvements of integrated circuits over the last decades. However, as device dimensions are within the range of a few atomic layers, present-day designers of integrated systems are facing an increasing number of issues and restrictions to meet the desired requirements of their designs. This includes, amongst others, delay as well as area restrictions but also the consideration of power consumption and lifetime reliability. Unfortunately, those mentioned design parameters are highly correlated so that the improvement of one aspect generally leads to deterioration of another, and vice versa. Nonetheless, the combination of adequate design techniques promises to offset the negative influences on one another. Thus, this paper proposes such a combined approach to tackle the problem of decreasing lifetime reliability in CMOS nanometer technologies. Such technologies, with device dimensions in the range of a few nanometers, suffer from an increased susceptibility to different kinds of failures during operation [1]. In contrast to previous technology generations, solutions within the manufacturing process are not sufficient anymore to deal with these kinds of issues. Accordingly, reliability concerns are not only an issue of manufacturing anymore, but also have to be considered in all abstraction layers of the design process. Thereby, three main strategies can be identified: (I) design techniques that detect errors [2][3], (II) techniques that detect and correct errors [4] and (III) those techniques that try to avoid or at least delay errors [5][6]. As those techniques of strategy (I) require another mechanism to cope with the detected error, they do not increase the reliability of the designs as aimed at in this work. Instead, our proposed approach in this contribution relates to strategy (III) and combines *Sleep Transistors* with the idea of modular redundancy to extend lifetime reliability of integrated circuits. Thus, techniques of strategy (II) can complement our approach so as to cope with errors as they can finally still occur.

The remainder of this contribution is organized as follows: section 2 presents the essential

fundamentals to ease the understanding of the following sections. The next section 3 introduces the proposed approach while section 4 presents an extended flow in order to quantify lifetime reliability of integrated circuits. The subsequent section 5 presents and discusses simulation results before section 6 finally concludes this contribution.

2 FUNDAMENTALS

The following subsections cover a brief overview of the necessary fundamentals of reliability, failure causes in nanotechnology, sleep transistors and the impact of redundancy.

2.1 Basics of reliability

The analytical term Reliability *R(t)* is to be understood as the probability of a system to perform as desired until time instance *t*. For example, $R(t_x) = 0.85$ states that there is an 85 % chance that the system is still running at time t_x. Moreover, the failure rate λ expresses the probability that a system fails in a given time interval. To clarify this definition, assume that 12 out of 100 systems fail in a given year. Accordingly, an individual system will fail with a probability of 12 % in that same year. For most cases, a constant failure rate λ is assumed over the useful system lifetime [7] so that reliability *R(t)* can be expressed by the exponential function, see Equation (1). In addition, closely related to the rather probabilistic expressions of reliability and failure rate is the Mean Time To Failure (MTTF), which is the average time that a system operates until it fails. Hence, it is equal to the expected lifetime and is expressed as the inverse of the constant failure rate λ, as shown in Equation (2).

$$R(t) = e^{-\lambda t} \tag{1}$$

$$MTTF = \int_0^\infty R(t)dt = \frac{1}{\lambda} \tag{2}$$

Even though, these equations are used in most calculations, it needs to be noted that the assumption of a constant failure rate λ is only valid for the regular lifetime. Thus, infant mortality as well as

wear-out mechanisms are excluded and are for the most part described by Weibull distributions [7]. Since corresponding measures are generally taken (such as burn-in right after manufacturing), we concentrate on the lifetime reliability.

2.2 Failure causes in nanotechnology

Integrated systems manufactured in nanotechnology suffer from an increasing number of physical failure causes during their lifetime. The recently most often reported ones are Electromigration (EM), Time Dependent Dielectric Breakdown (TDDB) and Negative Bias Temperature Instability (NBTI) [8]. Among these stated ones, the best-known failure mechanism is electromigration, which mainly concerns aluminum and copper interconnects with high current densities. Electromigration denotes the transport of material along an interconnect due to the gradual movement of the ions in a conductor caused by the electric current [8]. Due to this migration, material can be depleted or accumulated at other instances. As a consequence, highly resistive connections or even abrupt disruption can be created. Conversely, undesired connections can originate between interconnects when material is accumulated. Equation (3) shows a widely used model for the MTTF due to electromigration based on Black's equation [9][10]:

$$MTTF_{EM} = A_{EM}\left(J - J_{crit}\right)^{-n} e^{\frac{E_a}{k_B T}} \tag{3}$$

Here, A_{EM} is an empirically determined constant, J denotes the current density within the interconnect, J_{crit} is the critical level of current density required for electromigration, E_a terms the activation energy, k_B is Boltzmann's constant, T denotes the absolute temperature in Kelvin and n is another empirical constant. Hence, it can be concluded that during the runtime of the system the main driving forces for electromigration are high currents (as found in power supply lines), which lead to high current densities and high temperatures.

Gate oxide breakdown describes the appearance of a conducting path between the gate on the one end and the substrate, source or drain on the other end [11]. The breakdown can be caused by sudden

events, such as electro-static discharge, or by the rather slow destruction over time known as Time Dependent Dielectric Breakdown (TDDB). The latter cause is due to an autocatalytic loop in which overlapping charge traps create a conducting path between gate and substrate (or source/drain). This again leads to an increased current flow and heat dissipation. Consequently, thermal damage occurs even earlier and more charge traps are created. This positive feedback loop results in an accelerated breakdown and finally in a defect transistor [12]. Following from experimental work performed at IBM [8][13] the mean time to failure due to TDDB can be modeled as:

$$MTTF_{TDDB} \propto \left(\frac{1}{V_{DD}}\right)^{a-bT} e^{\frac{X+\frac{Y}{T}+ZT}{kT}} \tag{4}$$

Here, V_{DD} denotes the supply voltage and a, b, X, Y as well as Z are fitting parameters. In [8] and [13] the following values were determined for the fitting parameters $X = 0.759$ eV, $Y = -66.8$ eVK, $C = -8.37\text{E}{-4}$ eV/K, a = 78 and b = -0.081. By investigating these parameters and Equation (4) it can be concluded that TDDB is primarily a function of the applied voltage level at the gate of a transistor and the temperature.

Negative Bias Temperature Instability (NBTI) is a failure mechanism that becomes firstly apparent by a decrease of circuit performance. NBTI is mainly observed in PMOS transistors since they commonly operate with negative gate-to-source voltage. This temperature-activated effect occurs when voltage stress is applied to the transistor gate. The consequence of NBTI is a significant increase of the threshold voltage V_{th} of the transistor. This in turn results in higher delays and leakage currents of the affected integrated circuits. The physical reasons for NBTI are hole trapping in pre-existent oxide traps after manufacturing and the creation of interface states [14]. Thereby, the interface trap generation $N_{it}(t)$, which leads to a linear increase of V_{th}, can be expressed as [15]:

$$N_{it}(t) = 1.16\sqrt{\frac{k_f N_0}{k_r}(D_H t)^{1/4}} \tag{5}$$

Here, the mobile diffusing species are assumed to be neutral H atoms. N_0 is the concentration of

initial interface defects, D_H is the corresponding diffusion coefficient, and k_f and k_r are constants for the rates of dissociation and self-annealing, respectively. When the device is in the recovery phase, k_f becomes zero, and k_r is unchanged. In summary, during system operation NBTI depends on temperature and the applied voltage level at the gate of the transistor.

2.3 Power gating with Sleep Transistors

The application of *Sleep Transistors* for power gating is one of the most effective methods in order to reduce standby leakage [16][17][18]. A sleep transistor is referred to as a high threshold voltage device implemented with either a PMOS or NMOS transistor, or both. These transistors are applied as switches in order to disconnect the power supply lines (VDD or GND) from the actual logic blocks (see Figure 1). Hence, the control signal *Sleep mode* disconnects the block during idle phases of the attached logic. Thereby, the sleep transistors create a virtual power (in case of the PMOS) and/or a virtual ground (in case of the NMOS). That means in theory that during standby there are no leakage currents within the gated logic block. However, it should be noted that even when all sleep transistors are switched off a small leakage current still exists. This is mainly caused by sub-threshold leakage of the sleep transistors and parasitic capacitances within the logic block itself [16].

In recent years several strategies regarding sizing, placement and activation of sleep transistors were published [16][17][18]. Consequently, a comprehensive study of these findings is outside the scope of this work and can be found in the according publications. In a nutshell, the application of sleep transistors is a common and widely used approach and can for instance be found in current high-end processors [19].

2.4 Types of redundancy

Triple Module Redundancy (TMR) is a technique that allows increasing the reliability of integrated designs. Thereby, a module (that corresponds to a logic block) is implemented three times so that the same set of data can be executed in parallel. In case of different results at the outputs of the three

modules, a subsequent voter chooses the one which is most probably the correct one. Thus, the simplest approach is to mask the result of the module that differs from the other two. Under the assumption of an ideal voter (i.e. faultless) and a constant failure rate for all three logic blocks, the reliability of a design implemented with TMR R_{TMR} can be expressed by:

$$R_{TMR}(t) = 3R_{mod}(t)^2 - 2R_{mod}(t)^3 \tag{6}$$

Here R_{mod} is the reliability of a single module (i.e. a single logic block). With respect to Equation (2) the MTTF of the very same design is:

$$MTTF_{TMR} = \int_0^\infty \left(3e^{-2\lambda_{mod}t} - 2e^{-3\lambda_{mod}t}\right)dt = \frac{5}{6}MTTF_{mod} \tag{7}$$

Here λ_{mod} is the failure rate of a single module, and $MTTF_{mod}$ is the corresponding mean time to failure. It can be observed that TMR reduces the mean time to failure and, thus, cannot be considered as a technique to increase the expected lifetime of an integrated system. This perplexing result is due to the larger system size (more than three times) and thus a higher absolute failure rate over the entire lifetime of the design.

Another approach to exploit redundancy is Parallel Module Redundancy (PMR). Here, the design is implemented with several instances of the actual module. In contrast to TMR though, the voter has to be capable to detect which module produces a wrong result. Otherwise, the erroneous module cannot be masked when only two working modules remain. Hence, while this approach already works with two redundant instances, it requires a voter with additional functionality for error detection. Under these assumptions and an ideal voter, the MTTF of a design with PMR can be expressed by:

$$MTTF_{pmr} = \frac{1}{\lambda_{mod}} \sum_{i=1}^{N} i^{-1} = MTTF_{mod} \sum_{i=1}^{N} i^{-1} \tag{8}$$

Here, N is the number of redundant instances. A closer analysis reveals that a second instance increases $MTTF_{pmr}$ by 50 % while three instances result in a $MTTF_{pmr}$ which is just 83,3 % longer

compared to the reference design with just one module. While these results are better than with TMR, they necessitate error detection within the voter. Independent from the implementation of the voter though, both approaches (TMR and PMR) increase the power dissipation significantly by more than a factor of 3 (i.e. for three blocks within PMR).

In summary, the major drawbacks of module redundancy (including TMR and PMR) are the multiplication of area needs and power dissipation, as well as an increase of the overall delay due to the delay of the voter.

3 PROPOSED DESIGN APPROACH

The following two subsections introduce the essential idea of the proposed approach as well as requirements for the necessary control logic.

3.1 Essential idea

The important characteristic of power gating with sleep transistors is the fact that the gated logic can dynamically be disconnected from the power supply during the runtime of integrated circuits (see also section 2.3). Hence, within the disconnected state the gated logic is ideally without any inherent currents and voltages. While these are the key parameters under consideration in common publications, the idle logic block (i.e. no switching activity is present) also results in a reduction of local temperatures. Thus, during such an idle phase the three decisive parameters that reduce lifetime reliability of integrated circuits are eliminated or at least strongly reduced. As a consequence, the mean time to failure is prolonged approximately by the time that the design is in the idle phase.

One way to force longer idle times is to improve those related algorithms. Hence, if affected blocks can complete their tasks earlier, they can also stay longer in standby (i.e. in an idle phase). However, to speed up individual logic blocks is mostly not a viable option. Besides, in designs with high activity (as in streaming applications, medical environments etc.) the potential to extend the idle

times is strongly limited [18]. Therefore, our proposed approach is based on a different idea to extend the idle time of a gated logic block: namely, on a redundant implementation. Thereby, each gated module (i.e. each logic block) is implemented at least two times, see also Figure 2. During the runtime though, only one of these instances is active while the others are disconnected from the power supply. Consequently, the resulting mean time to failure of the proposed design with sleep transistors and redundant modules $MTTF_{STmr}$ can be expressed by:

$$MTTF_{STmr} = \sum_{i=1}^{N} MTTF_{mod} = N \cdot MTTF_{mod} \tag{9}$$

Where N is the number of redundant instances. Equation (9) refers to the ideal case where any additional logic is neglected and the gated modules are completely disconnected from the power supply. Under these eased assumptions, two redundant and gated instances double the MTTF of the design. Hence, compared to an ideal implementation of Parallel Module Redundancy (PMR) with two redundant instances, our proposed approach offers an MTTF that is 33 % higher. In case of four redundant instances, the improvement compared to PMR even increases to 92 %, while it improves by 400 % compared to the initial design with no redundancy.

It should be noted that the proposed approach is not limited to designs with high activity. In other words, the MTTF can also be reduced in designs with long idle times, where all blocks are disconnected from the power supply. Then the resulting equation for the MTTF under ideal conditions becomes:

$$MTTF_{STmr,Ia} = N \cdot \left(\frac{MTTF_{mod}}{1 - p_{standby}} \right) \tag{10}$$

Where $p_{standby}$ is the percentaged amount of time that the system is in an idle state (i.e. in standby).

An additional advantage of the proposed approach is the very small increase of power dissipation. This follows from the fact that at a given time at most one instance is active while the necessary overhead for additional logic is also comparatively small (see next subsection).

3.2 Additional logic

In order to properly work, additional logic is required to multiplex the results from the currently active module to the subsequent module. This is implemented by multiplexers that are placed behind the redundant instances. Figure 2 shows an illustration of the proposed approach and the position of the described multiplexers. Here, a simple 2:1 multiplexer is shown to forward the correct signals from the redundant instances to the subsequent module 2.

Commonly, power gated logic requires additional clock cycles before the logic can be fully operated again (i.e. wake-up time [18]). Hence, it is not feasible to connect the signals controlling the sleep transistors (here Sleep1 and Sleep2) also directly to the multiplexers. Instead, a clock scheme as given in Figure 3 should be applied to ensure data consistency. The figure shows the control signals for the sleep transistors with respect to Figure 2 (Sleep1, Sleep2) and the control signals for the multiplexer (MUX-ctrl and its inverse). At time t = 250 ns, instance 2 is active while instance 1 is in sleep mode (consider the low-active control signals). Then at time t = 355 ns, instance 1 is turned on and starts to return into an active state. However, the multiplexer still forwards the outputs of instance 2. At time t = 375 ns, instance 1 is fully active and the multiplexer starts to forward the outputs of instance 1. Just slightly after the multiplexer was switched, instance 2 goes into standby and is gated off the power supply. Accordingly, the appropriate scheme is applied when instance 2 returns to its active state and instance 1 goes back to the sleep mode.

For a comprehensive investigation, it has to be considered that the lifetime of the system also depends on the MTTF of the multiplexers. The multiplexers though are realized as transmission gates (see Figure 2), whereas only one path is active at a time. Thus, the impact of failure mechanisms (like electromigration, TDDB and NTBI) are also correspondingly smaller. Nevertheless, it is reasonable to apply special design strategies for the multiplexers as well, like transistors with thicker gate oxide [20] and wider wires.

4 PROPOSED FLOW TO QUANTIFY RELIABILITY

The simulation of physical mechanisms and their impact on the characteristics of integrated circuits is a general problem. In particular, there exists no common procedure in order to quantify measures of reliability. Such a task becomes even more complicated when time-dependent failure mechanisms shall be taken into consideration. However, we propose a potential flow to model such failure mechanisms and their impact on circuit characteristics.

4.1 Types of modeling of failure mechanisms

Several models for individual failure mechanisms within integrated circuits can be found in the literature [7][8][10], whereas SPICE simulations are reported as the most accurate approach used by circuit designers. The accuracy is also the reason why we selected transistor-level analysis for our approach, although it can necessitate major computational efforts and simulation times, which limits the maximum number of elements within an investigation.

Those existing models for reliability simulations based on transistor-level can be categorized into (1) models based on electronic components (i.e. transistors, resistors, etc.) and (2) models based on analytical equations. Examples for category (1) can be found in [21] where TDDB is modeled with transistors and resistors and in [8] where electromigration is modeled with resistors and capacitors. The second category includes transistor models with aging parameters that are extracted from tools like Cadence BSIMPro+ or as described in [22]. Another solution is the application of voltage controlled voltage sources (VSCS) or current sources (VCCS) to model transistor behavior [23].

All above listed approaches simulate the circuit behavior at a certain time instance, e.g. after 7 years of runtime. Thus, it is possible to analyze the design performance after a given time but not its variation during the aging process. However, particularly for the quantification of the mean time to failure of a circuit it is important to monitor system behavior over time. Therefore, our proposal is the

extension of the models applied in both previously described categories by controllable elements which can dynamically be adapted during simulation. For the following introduction, the term device denotes a part of the actual design, for instance a transistor or a capacitor. In contrast, the term element is related to parts of the device model and can be both electronic components and their parameters.

We identified three possible techniques for the implementation of adaptable elements: (I) voltage or current controlled active/passive components, e.g. voltage controlled voltage sources (VCVS) or voltage controlled resistors (VCR), (II) variables that are used in the algebraic expressions of device parameters, (III) extended models with VerilogA or VHDL-AMS descriptions. The first technique requires that the time-dependent behavior of reliability can be emulated by an inherent voltage/current source or an inherent passive element, respectively. Then, during simulation, the value of this element is varied via a voltage or current source (see example in Figure 4).

The second technique demands the identification of device parameters which suffer under time-dependent failure mechanisms. That can be the threshold voltage of a transistor or the resistance of an interconnection. During simulation, the identified parameter is varied via a current or voltage source (see Figure 5). The third technique for the realization of adaptable elements is comparable to both previous ones. The difference is that the element or even the whole device is realized in a hardware description language, which comprehends analog and mixed-signal extensions, e.g. VerilogA or VHDL-AMS. These languages allow a more complex emulation of the behavior of the circuit device. However, this solution can increase the simulation time. As in both former techniques a voltage or current sources is used to adjust the speed of device degradation (see Figure 6) [24].

4.2 Flow to detect errors and quantify reliability

The basic architecture of the test environment is similar to common mixed-signal structures. That is, the input signals of the Design Under Test (DUT) are generated by a digital circuit written in VHDL-

AMS or VerilogA. Similarly, the outputs of the DUT are also connected to a digital control block, which tests the correctness of the results and monitors the time of the first erroneous result. The devices of the DUT are based on adaptable models as described in the preceding subsection. During simulation, the degradation of the devices is controlled via a voltage source. In order to model TDDB accurately, the signal probabilities of the inputs of the devices are also examined and assigned to the according models. Thereby, we consider the fact that TDDB depends on the voltage level at the gate (see section 2.2). For instance, if no voltage is supplied to the gate of an NMOS (gate voltage $V_g = 0$), there is no electric field and TDDB does not degrade. On the other hand, if V_g = VDD an electric field and TDDB are present and the $MTTF_{TDDB}$ is smaller compared to the general case with equal signal probability. The same applies for PMOS, but with inverse voltage levels − the critical case is $V_g = 0$. Besides, we also use the examined signal probabilities (which comprehends switching activity) to estimate the related currents. These are assigned to the according models for electromigration (see section 2.2). Lastly, such examined activity and probability values are of similar interest for NBTI and are used correspondingly (see section 2.2).

Hence, the proposed flow is as follows:

1) Synthesis of the design that is to be analyzed
2) Generation of input signal patterns and control blocks for the outputs of the DUT
3) Extraction of maximum clock frequency, signal probabilities and switching activities
4) Conversion into a SPICE netlist whereas the used devices are based on the adaptable models
5) Insertion of degradation factors which are derived from signal probabilities and activities
6) Insertion of pulsed voltage sources with slopes that mimic the MTTFs of the individual devices
7) Simulation of the DUT until the first erroneous output occurs. Those simulations are repeated with varying and random start values.

5 RESULTS OF THE SIMULATIONS

In this section, the setup of the test environment is presented before the obtained simulation results are discussed.

5.1 Setup of the test environment

All simulation results are based on designs from the ISCAS benchmark suite [25], which we implemented in a predictive 22 nm technology [26]. The considered degradation of devices comprehends at this stage models for TDDB and electromigration. That means that each transistor of the used SPICE netlists is replaced by the model for TDDB as depicted in Figure 4. Hence, the time-dependent degradation is realized by the voltage controlled resistor (VCR) and the according definition of the pulsed voltage source. Likewise, electromigration is modeled with a VCR and corresponding voltage source at the output of each logic gate. We are aware that the used models are fairly simple and allow reliability predictions with limited accuracy. However, the results of the simulations clearly depict the characteristics of different implementations and allow a relative evaluation among one another. Furthermore, the current setup confirms the proposed flow and can also be used with more complex and thus highly accurate models.

As shown in Figure 2, the multiplexers are implemented with transmission gates and the control signals as well as their inverse are generated by a simple control unit. Furthermore, all digital blocks (i.e. input signal generation and the control blocks at the outputs of the DUT) are implemented in VerilogA. The automated generation of these blocks, the duplication of module instances, the insertion of sleep transistors and multiplexers and the estimation of the clock frequency is done by a tool specifically written for these tasks.

5.2 Results and discussion

As a first step and a reference point, each design was simulated as the initial version without any redundant blocks. Subsequently, those designs were modified according to the proposed design

approach in section 3. Hence, each design was duplicated and complemented with the multiplexers and the control logic. Within each redundant instance, the logic gates were grouped in order to connect the sleep transistors. The applied clock scheme for the sleep transistors and the multiplexers follows the described scheme in Figure 3. Thereby, the wake-up time was set to one clock cycle which refers to the small groups of logic gates attached to one sleep transistor. Lastly, we adjusted the slope of the pulsed voltage sources so as to mimic an expected lifetime of an individual device of around 300 clock cycles (see step 6 in subsection 4.2). The simulations were then repeated 100 times for each design with varying start values in order to calculate average values.

The first set of results is shown in Figure 7 where the improvements of the MTTF are plotted for several designs from the ISCAS benchmark suite. Accordingly, the MTTF of the designs with our proposed approach improve on average by an impressive 280 % with reference to the initial designs without redundancy. This is even higher than expected, which can be explained by two reasons. On the one hand, the delay and the maximum output voltage of a logic cell do not degrade linearly under the presence of TDDB. Hence, in this case the extension of the active time by a certain factor reduces the deterioration by a slightly higher factor [26]. On the other hand, the clock frequency differs for the versions with and without redundancy. More precisely, the maximum clock frequency is lower within those redundant designs, which in turn relaxes the delay constraints within the logic blocks. In order to omit this factor, we additionally simulated all initial designs with the same clock frequency as their counterparts. The adjusted results for the MTTF are plotted in Figure 7 (see “w/ relaxed delay”). The relative MTTFs of our proposed approach are slightly smaller now, but still exhibit an impressive improvement of 216 % on average.

However, the improvements come at the price of redundancy and thus additional area needs. Such numbers are given in Figure 8 for the various ISCAS designs and in relation to the initial designs. According to that, the area needs are roughly doubled for all designs because of the duplicated

modules. The additional logic (see subsection 3.2) only has minor impact on the overall area needs. To clarify this fact, the numbers of required multiplexers are also depicted in Figure 8 in relation to the overall number of logic gates within the modules themselves. Hence, the small percentages (average of 2.2 %) mean that the designs have only few output signals in comparison to their internal number of logic gates.

Even though there are quite few multiplexers in the designs, they prolong the critical paths and thus increase the delay. This is quantitatively illustrated in Figure 9, where an average delay increase of 2.3 % compared to the initial designs can be observed. The delay increase is largest for the design c432 because it is also the smallest. Contrariwise, the delay penalty becomes smaller the larger the design is. Or more precisely, the larger the critical path is within the duplicated modules.

Lastly, the power dissipation is considered in Figure 10 for the dynamic power and in Figure 11 for the leakage power. The designs here exhibit an average increase of dynamic power by 5.5 %, which is mainly attributed to the additional logic (e.g. the multiplexers) and the parasitic capacitances due to more complex wiring. The increase is fairly small though when considering that the area is more than doubled (see Figure 8). The reason is that only one module is active while the second is in sleep mode, which clearly favors the proposed approach over TMR and PMR where all redundant modules operate in parallel (i.e. two and three times the power dissipation, respectively). In case of the leakage power in Figure 11, the doubled area is the decisive factor again. Therefore, the leakage power also doubles since the reference designs are equipped with sleep transistors too to achieve a fair comparison.

In summary, the proposed approach can significantly increase the MTTF while the additional area needs do not offset those improvements. Furthermore, the slight penalty of the dynamic power consumption is negligible and clearly superior to common redundancy approaches such as TMR and PMR. Hence, the proposed approach is a power-efficient technique to enhance the reliability of

integrated circuits.

6 CONCLUSION

Integrated circuits in nanometer technology are continuously more susceptible to severe failure mechanisms. This alarming development necessitates design techniques to improve the lifetime reliability. Hence, the presented work proposes an approach that combines the ideas of sleep transistors and modular redundancy in a beneficial way. Thereby, the approach aims at increased lifetime reliability while the impact on delay and power dissipation is kept to a minimum. In order to quantify reliability, we also proposed an extended flow that allows analyzing circuit designs under the impact of time-dependent deterioration. Finally, the simulation results show that the proposed design approach can improve the Mean Time To Failure (MTTF) by 216 % on average, while the dynamic power and the delay slightly impair. However, those results clearly outperform such common approaches of modular redundancy and allow a power-efficient enhancement of reliability.

Further investigations concentrate on the application of more complex failure models and on algorithms to selectively implement redundancy. Such algorithms promise to adaptively trade area and delay penalties with improvements of the MTTF [20].

REFERENCES

[1] J. Srinivasan, S. Adve, P. Bose, and J. Rivers, "The impact of technology scaling on lifetime reliability", Proceedings of IEEE International Conference on Dependable Systems and Networks, (**2004**).

[2] P. Bernardi, L. M. V. Bolzani, M. Rebaudengo, M. S. Reorda, F. L. Vargas, and M. Violante, "A new hybrid fault detection technique for Systems-on-a-Chip", IEEE Transaction on Computers, (**2006**). 55, 2, pp. 185-198.

[3] R. Datta, A. A Jacob, A. U. Diril, A, Chatterjee and K. Nowka, "Adaptive design for performance-optimized robustness", Proceedings of International Symposium on Defect and Fault-Tolerance in VLSI Systems, (**2006**), pp. 3-11.

[4] S. Mitra, N. Seifert, M. Zhang, Q. Shi, and K. S. Kim, "Robust System Design with Built-In Soft-Error Resilience," Computer, (**2005**), Vol. 38, Issue 2, pp. 43-52.

[5] T. Inukai, T. Hiramoto, and T. Sakurai, "Variable threshold CMOS (VTCMOS) in series connected circuits" Proceedings of the International Symposium on Low Power Electronics and Design, (**2001**), pp. 201-206.

[6] J. Tschanz et al., "Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage", IEEE Journal of Solid-States Circuits, (**2002**), vol. 37, pp. 1396-1402.

[7] I. Koren and C. Krisha, Fault-tolerant systems, Morgan Kaufmann (**2007**).

[8] J. Srinivasan, S. V. Adve, P. Bose, J. Rivers, and C.-K. Hu, "RAMP: A Model for Reliability Aware Microprocessor Design", IBM Research Report, RC23048 (**2003**).

[9] J. R. Black, "A brief survey of electromigration and some recent results", In IEEE Transactions on Electron Devices, (**1969**), pp. 338-347.

[10] "Failure Mechanisms and Models for Semiconductor Devices", JEDEC Publication JEP122-A, Jedec Solid State Technology Association (**2002**).

[11] J. Stathis, "Reliability limits for the gate insulator in cmos technology", IBM Journal of Research & Develop, (**2002**), Vol. 46, Nº 2/3, pp. 265-286.

[12] D. Crook, "Method of determining reliability screens for time dependent reliability breakdown", IRPS, (**1979**).

[13] E. Wu, J. Suñé, W. Lai, E. Nowak, J. McKenna, A. Vayshenker, and D. Harmon, "Interplay of voltage and temperature acceleration of oxide breakdown for ultra-thin gate dioxides", Solid-state Electronics Journal, (**2002**), Vol. 46, pp. 1787-1798.

[14] T. Grasser and B. Kaczer, "Evidence that two tightly coupled mechanisms are responsible for negative bias instability in oxynitride MOSFETs", (**2009**), IEEE Transactions on Electronic Devices, Vol. 56, Nº 5, pp. 1056–1062.

[15] E. Maricau and G. Gielen, "NBTI model for analogue IC reliability simulation", Electronics Letters, (**2010**), Vol. 46, Nº 18.

[16] M. Powell, S.-H Yang, B. Falsafi, K. Roy, and T. N. Vijaykumar, "Gated-Vdd: A circuit technique to reduce leakage in deep-submicron cache memories", Proceedings of International Symposium on Low Power Electronics and Design, (**2000**), pp. 90-95.

[17] A. Ramalingam, B. Zhang, A. Davgan, and D. Pan, "Sleep transistor sizing using timing criticality and temporal currents", Proceedings of Asia South Pacific Design Automation Conference, (**2005**), pp. 1094-1097.

[18] K. Shi and D. Howard, "Challenges in sleep transistor design and implementation in low-power designs", Proceedings of Design Automation Conference, (**2006**), pp. 113.

[19] “Designing for power – Intel leadership in power efficient silicon and system design” www.intel.com/technology (**2004**).

[20] C. Cornelius, F. Sill, H. Sämrow, J. Salzmann, D. Timmermann, and D. da Silva Jr., “Encountering gate oxide breakdown with shadow transistors to increase reliability“, Proceedings of Symposium on Integrated Circuits and Systems Design, (**2008**), pp. 111-116.

[21] M. Renovell, J. Gallière, F. Azaïs and Y. Bertrand, “Modeling the random parameters effects in a non-split model of gate oxide short”, Journal Electronic Testing, (**2003**), Vol. 19, Nº. 4.

[22] R. Vattikonda, W. Wang, and Y. Cao., “Modeling and minimization of PMOS NBTI effect for robust nanometer design”, Proceedings of the Design Automation Conference, (**2006**), pp. 1047-1052.

[23] H. Li and Y. Chen, “An overview of non-volatile memory technology and the implication for tools and architectures,” Proceedings of Design, Automation & Test in Europe, (**2009**), pp. 731-736.

[24] M. Kole, "Circuit reliability simulation based on Verilog-A", Proceedings of IEEE Behavioral Modeling and Simulation Workshop, (**2007**), pp.58-63.

[25] M. Hansen, H. Yalcin, and J. P. Hayes, “Unveiling the ISCAS-85 Benchmarks: A Case Study in Reverse Engineering”, IEEE Design & Test, (**1999**), Vol. 16, Nº 3, pp. 72-80.

[26] W. Zhao, and Y. Cao, "New generation of Predictive Technology Model for sub-45nm early design exploration," IEEE Transactions on Electron Devices, (**2006**), Vol. 53, Nº. 11, pp. 2816-2823.

[27] M. Renovell, J.M. Galliere, F. Azais, and Y. Bertrand, "Delay Testing of MOS Transistor with Gate Oxide Short," Proceedings of Asian Test Symposium, (**2003**), pp.168-173.

FIGURES AND TABLES

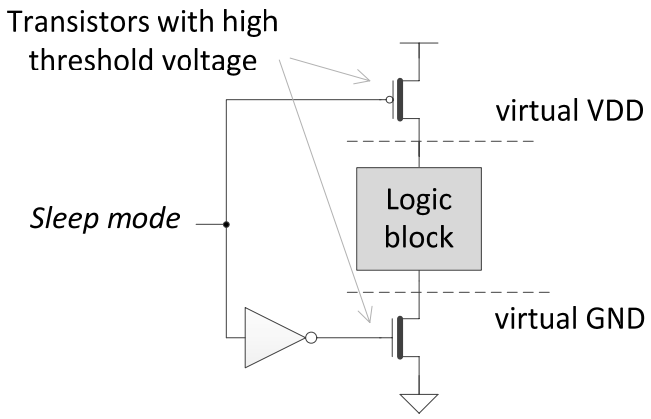

Figure 1. Illustration of the fundamental approach to reduce standby leakage within a logic block with sleep transistors

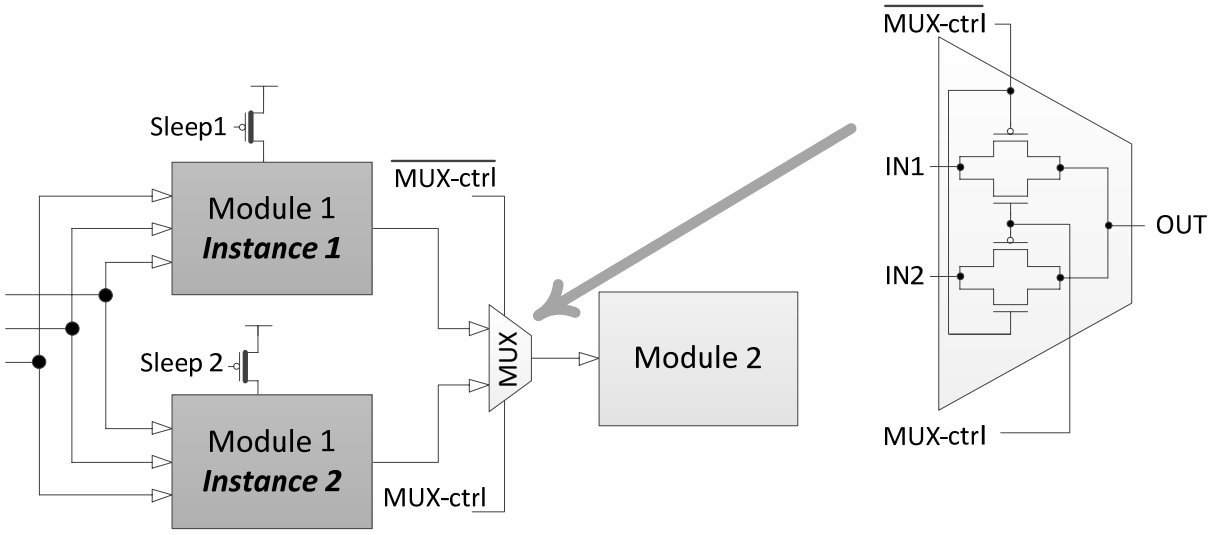

Figure 2. The proposed approach with two redundant instances, whereas the results of the active instance are forwarded by the subsequent multiplexer (implemented with transmission gates)

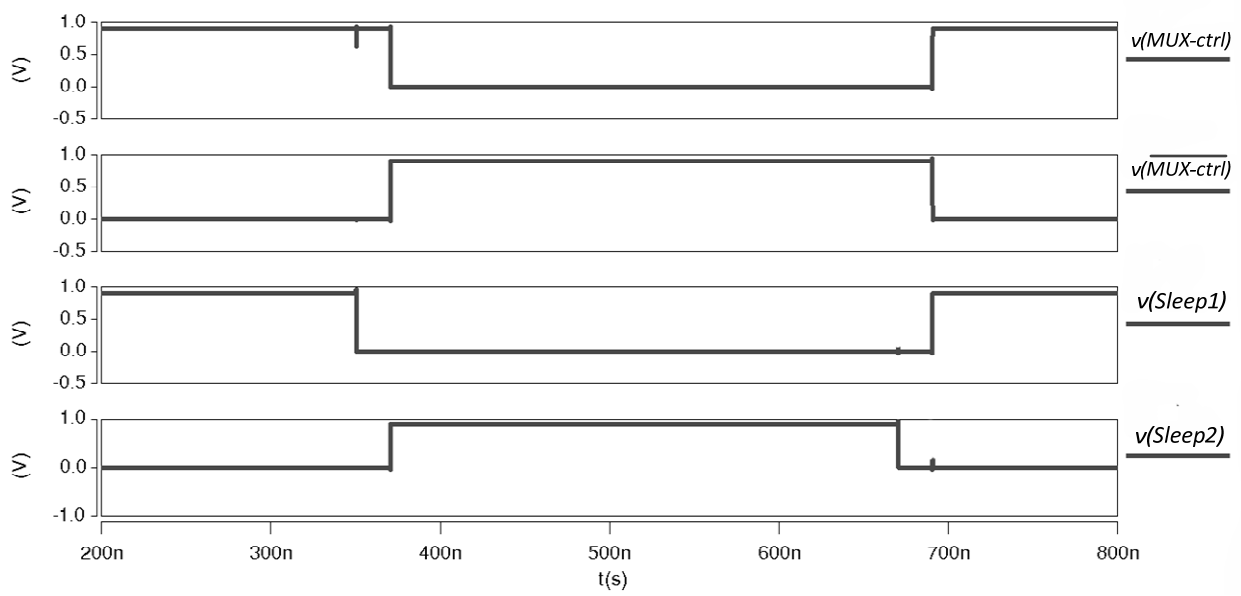

Figure 3. Timing scheme for the control signals of the multiplexers and the sleep mode in order to account for the wake-up time after an idle phase of a module

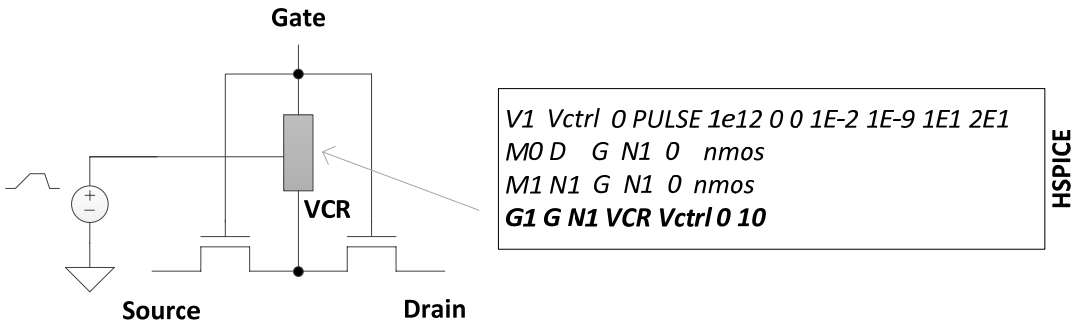

Figure 4. Example (incl. HSPICE code) for the modeling of TDDB with a voltage controlled resistor (VCR)

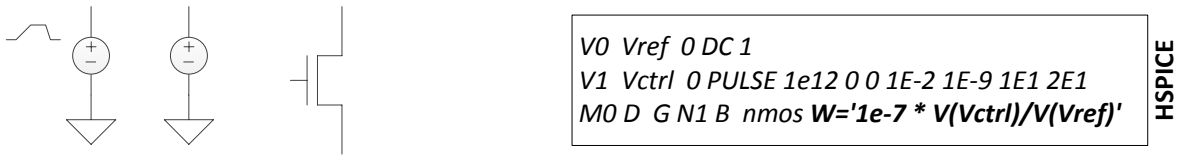

Figure 5. Example (incl. HSPICE code) for the modeling of varying gate width based on the algebraic expression

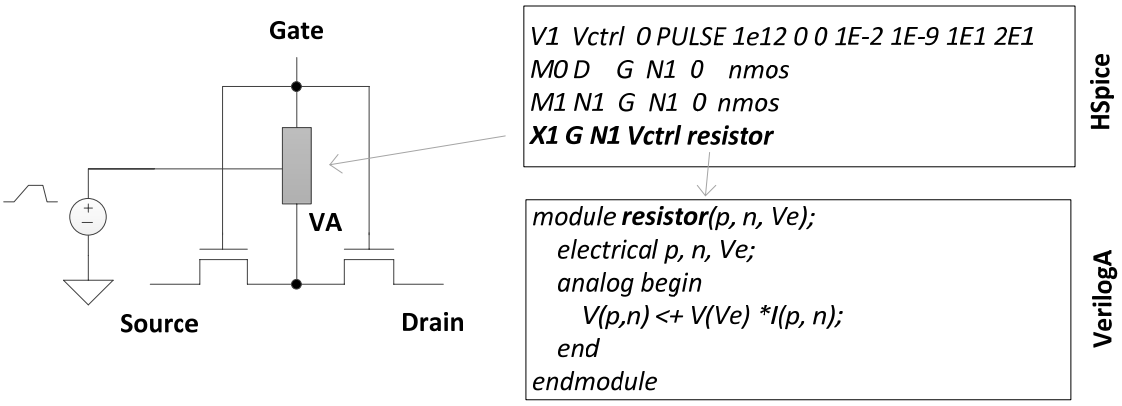

Figure 6. Example (incl. HSPICE and VerilogA code) for the modeling of TDDB with a resistor described in VerilogA

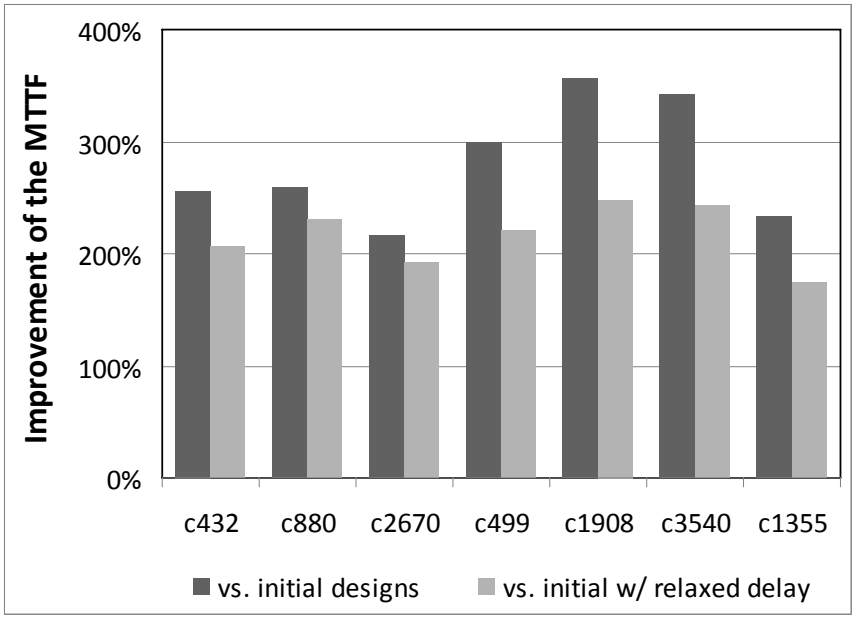

Figure 7. Improvement of the MTTF for the proposed approach and several ISCAS designs in comparison to initial designs with and without relaxed delay constraints

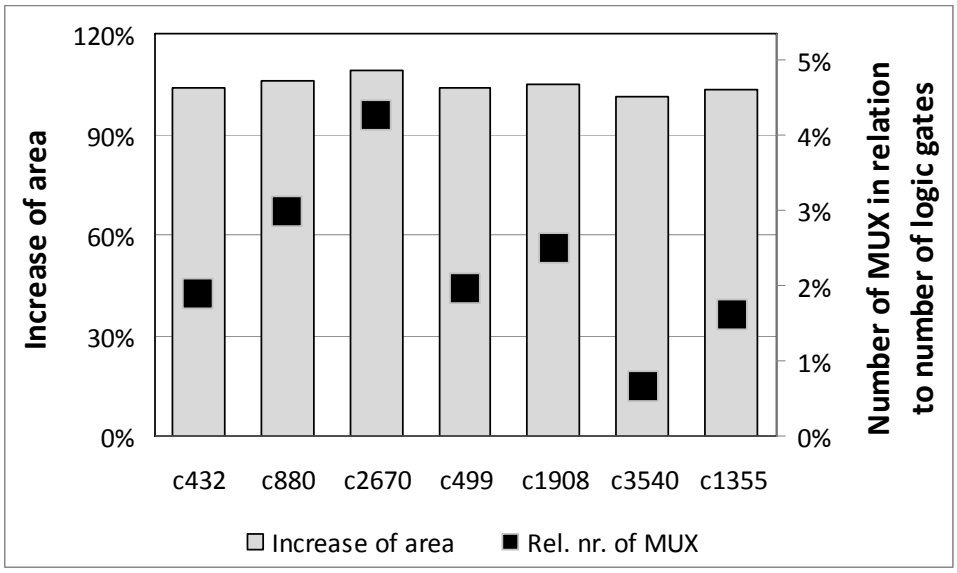

Figure 8. Area needs increase by about 100 % due to the duplicated modules, this is because of only few additional MUX in relation to the number of logic gates within the designs themselves

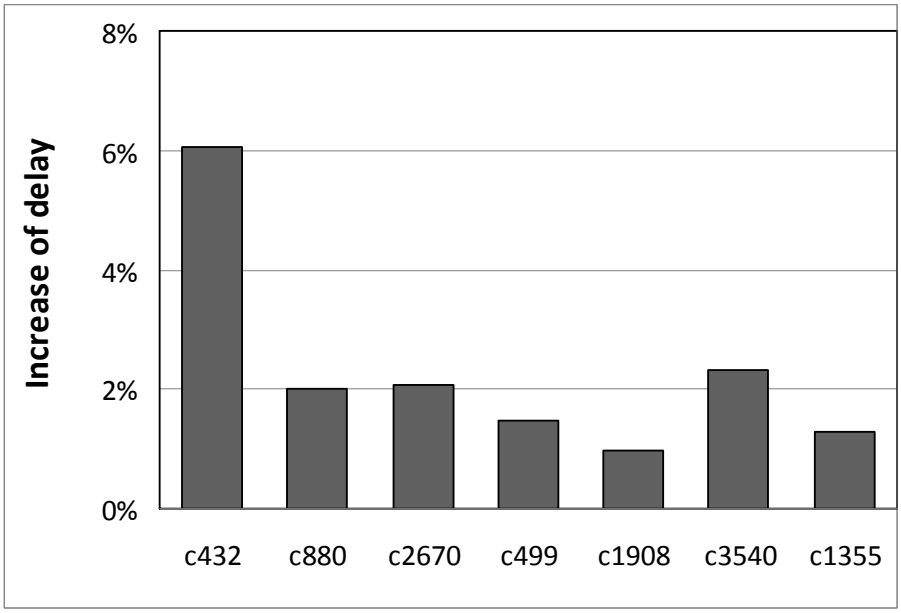

Figure 9. Those designs with redundancy exhibit a slight delay penalty due to the multiplexers

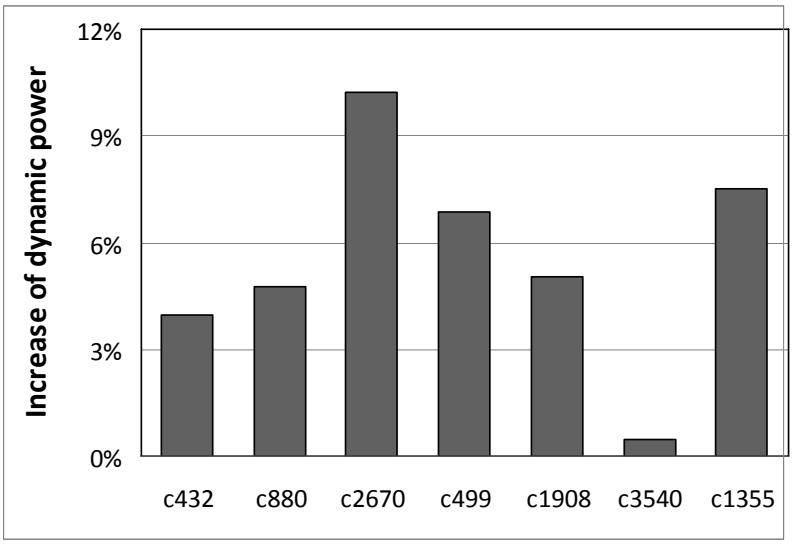

Figure 10. Increase of dynamic power for the ISCAS designs with redundancy compared to the initial designs

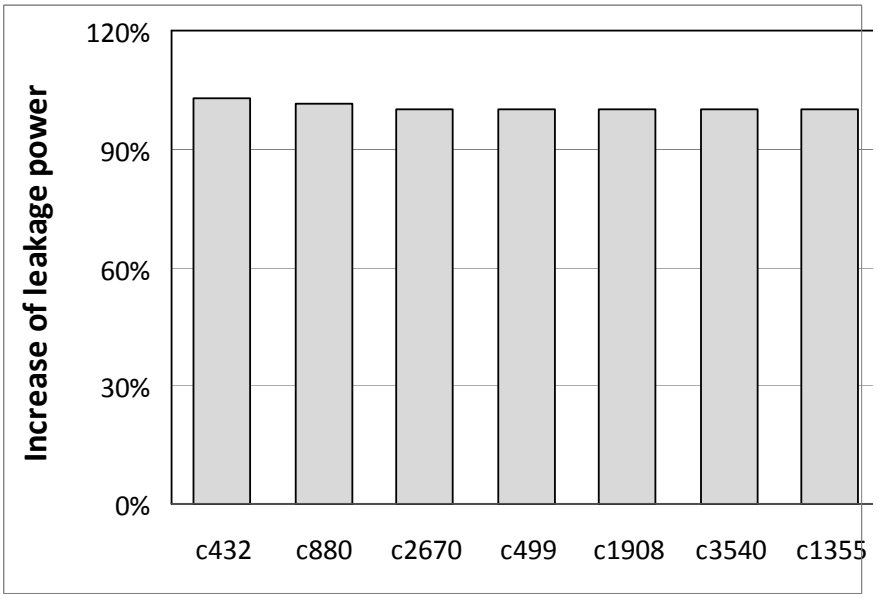

Figure 11. Leakage power is roughly doubled for all redundant design due to the duplicated logic instances

BIOGRAPHIES

Claas Cornelius studied information and electrical engineering at the University of Rostock, Germany, and graduated as Dipl-Ing. in 2005. Since then, he has been scientific co-worker at the Institute of Applied Microelectronics and Computer Engineering. He received a Ph.D. (Dr.-Ing.) degree in 2011 for his works on complex integrated systems, and he coordinates activities within the Network-On-Chip Research Group. His research interests comprehend reliability, low-power, dynamic circuit techniques and communication-centric systems.

Frank Sill Torres was born in Germany. He obtained the diploma in electrical engineering in 2002 and the Ph.D. degree in electrical engineering in 2007 from the University of Rostock, Germany. Until 2010 he worked as research associate at the Department of Electrical Engineering, Federal University of Minas Gerais, Brazil, in the field of integrated analog-digital converter as well as reliable and robust design. He is now adjunct professor at the Department of Electronic Engineering, Federal University of Minas Gerais, Brazil, where his main research interests involve design for reliability, low-leakage techniques and VLSI design for nanoscale technologies.

Dirk Timmermann studied electrical engineering at the University of Dortmund and graduated as Dipl.-Ing. in 1984. Afterwards, he worked until 1989 as scientific co-worker at the Fraunhofer Institute of Microelectronic Circuits and Systems. Under Prof. Dr. Hosticka's guidance (Duisburg) he worked in the department of signal processing and system design. In July 1990 Dirk Timmermann received his Ph.D. (Dr.-Ing.) from the Department of Electrical Engineering at the University of Duisburg and worked subsequently as scientific co-worker and project-leader at the Fraunhofer Institute of Microelectronic Circuits and Systems. In 1993 he became Professor for Computer Engineering at the University of Paderborn. Since 1994 he is a full University Professor for

Computer Engineering (School for Informatics and Electrical Engineering) at the University of Rostock. Concurrently, he is the managing director of the Institute for Applied Microelectronics and Computer Engineering. His research interests involve digital CMOS-circuits and systems, low power systems, sensor networks and self-organizing systems, wireless and wired communication.