Comparison of Strategies for Redundancy to improve Reliability concerning Gate Oxide Breakdown

Hagen Saemrow, Claas Cornelius, Frank Sill, Andreas Tockhorn, and Dirk Timmermann

***Abstract*—Integrated circuit designers are facing increased reliability and yield concerns due to the scaling of device dimensions towards atomic scales. Thus, lots of research efforts have focused on soft-errors and system-level approaches. However, only very little effort has been put into low-level solutions to enhance lifetime reliability. Therefore, this paper presents different design techniques that apply redundancy to improve reliability as regards gate oxide breakdown. Based on a Wallace multiplier, we compare three design strategies on different abstraction levels and demonstrate significant reliability improvements, especially when applied on transistor level. The achieved results show that lifetime reliability (in respect of oxide breakdown) increases up to more than 2500 % at constant delay, but at the price of overhead for area as well as power consumption. It should also be noted that the presented strategies can additionally improve defect yield (e.g. due to stuck-open transistors).**

***Index Terms*—Integrated circuit design, Redundant systems, Reliability, Gate oxide breakdown**

I. Introduction

Aggressive scaling of integrated circuits has resulted in ever increasing performance in the past decades. On the other hand, as we are approaching the limits of nanotechnology reliability and power concerns arise. For instance, design and process error margins due to material defects and imperfections are reduced which has direct impact on product yield. Furthermore, miniaturization leads to an increased sensitivity of transistors and interconnects to different kinds of failures during system operation. Hence, issues of and solutions for lifetime reliability will have to be investigated thoroughly. What makes the situation even more severe is that reliability issues, like Time-Dependent Dielectric Breakdown (TDDB), electromigration or thermal cycling, worsen with non-ideal scaling, increased transistor count and power density as well as adaptive processing [1].

Manuscript received November 7, 2008.

H. Saemrow, is with the Department of Electrical Engineering, University of Rostock, Germany. (phone: +49-381-498-7278; fax: 49-381-498-1187251; e-mail: hagen.saemrow@uni-rostock.de).

C. Cornelius, is with the Department of Electrical Engineering, University of Rostock, Germany. (e-mail: claas.cornelius@uni-rostock.de).

F. Sill is with the Department of Electrical Engineering, Federal University of Minas Gerais, Brazil (e-mail: franksill@ufmg.br).

A. Tockhorn, is with the Department of Electrical Engineering, University of Rostock, Germany. (e-mail: andreas.tockhorn@uni-rostock.de).

D. Timmermann, is with the Department of Electrical Engineering, University of Rostock, Germany. (e-mail: dirk.timmermann@uni-rostock.de).

One of the fundamental components for reliability, performance and power consumption is the gate oxide which is the dielectric isolation between the transistor input and the conducting channel. The thickness of gate oxide comprises only a few atomic layers (<20 Å) in current technologies. Due to this fact, non-ideal scaling of the supply voltage and increased electric fields, the gate oxide has become highly vulnerable to breakdown mechanisms causing transistor defects and logical malfunctions.

The point of time a conducting path is generated between the gate and the substrate is called gate oxide breakdown [2] whereas the cause can originate from two different situations. Firstly, sudden damage occurs due to extreme overvoltage and leads to non-isolating gate oxide, e.g. caused by Electro-Static Discharge (ESD). Secondly, a rather slow destruction over time is also possible, called Time-Dependent Dielectric Breakdown (TDDB). Thereby, charge traps start to form in the gate oxide during operation which causes an autocatalytic loop of events: overlapping charge traps form a conducting path between gate and substrate which leads to increased current flow as well as heat dissipation. This again causes thermal damage and, hence, more charge traps. This positive feedback loop results in an accelerated breakdown and finally in a defect transistor [3][4]. In the course of oxide destruction, the breakdown is distinguished into soft and hard breakdown depending on whether the transistor still operates in some degree or not. This means that the hard breakdown is in most cases the final result of soft breakdown.

Because of the rising transistor count per die, system failures caused by gate oxide breakdown will increase the number of system failures with every new technology generation. Unfortunately, full functional system tests are not feasible due to the rising complexity of integrated systems. Hence, tool assisted insertion of reliability mechanisms into the design flow will be one of the key priorities in the future [5]. In the following, different strategies for inserting redundancy are presented which can easily be embedded in current CAD tools.

Section II describes related work that motivates the chosen approach. Subsequently, fundamentals of how to compute quantitative measures for reliability and the different design strategies are presented in section III. Lastly, achieved results for a Wallace multiplier with and without defects are shown and their implications discussed in section IV before section V concludes the paper.

II. Related Work

Up to now, mostly manufacturing engineers focused on reliability and yield enhancements, but against the background of nanotechnology circuit designers are becoming more aware of the capabilities for improvement during system design.

Low-level techniques which enhance defect yield are already established in circuit design. A common approach used in memory manufacturing is static reconfiguration. Thereby, defective parts are disconnected from and spare parts are connected to the system by using laser fuses [6]. Another way for increased yield is achieved by layout modifications which rearrange parts of the system to optimize the design concerning yield. Possible optimizations are for example via duplication, length-minimization and widening of wires or layer reassignment [7].

Recently, a lot of techniques for soft-error resilience have also been published. Soft-errors are caused by radiation events – like neutrons, α-particles or electromagnetic interference – which can change the amount of electrical charges on internal nodes. These changes can result in transient failures of logical values (mostly happening in memory and registers). Technical solutions vary from RC-filters and additional capacitances to harden internal vulnerable nodes [8] to approaches for error detection – as in [9] where debug resources are used to minimize the design overhead.

By contrast, little effort has been made so far on techniques to enhance lifetime reliability. One such high-level approach was published in [10] where a dynamic system management adapts the operating conditions in response to an observed hardware usage to stay within a given reliability target. A very different approach is made in [11] where it is assumed that device failures cannot be prevented but have to be resolved. Therefore, redundant transistors are inserted randomly into the design to increase yield as regards stuck-open transistors. This idea was extended in [12] where the redundant transistors where inserted only at those instances that are most vulnerable to TDDB which increases not just the yield but also lifetime reliability.

Based on the promising results for redundant transistors, this paper compares the results of redundancy for different levels of abstraction because of the difficulties to transfer the existing transistor level approaches to CAD tools and given gate libraries. The contribution will show the significant differences for the different levels of abstraction by presenting figures on reliability, delay and power consumption.

III. Approach and Setup of Investigation

A. Fundamentals of Reliability

The reliability $R_{SYS}(t)$ of a system represents the probability of the system to perform as desired until time t. When assuming a constant failure rate λ_{SYS} – which represents the rate at which an individual system suffers from individual faults – the following equation for reliability can be derived [13]:

$$R_{SYS}(t) = e^{-\lambda_{SYS} t} \quad (1)$$

Closely related to the probabilistic term for reliability is the Mean Time To Failure ($MTTF_{SYS}$) which is the average time a system operates until it fails. It is equal to the expected lifetime if the system cannot be repaired. It can be calculated by [13]:

$$MTTF_{SYS} = \int_0^\infty R_{SYS}(t)\, dt \quad (2)$$

When a complex system consists of a number of components with known values for reliability, the system reliability can be derived by structuring the components in parallel and serial segments. A system configuration with components in series leads to a system which fails if any component fails. Thus, the reliability $R_S(t)$ of a serial system with n equal components and constant failure rate λ is described by [14]:

$$R_S(t) = \left[R(t)\right]^n = \left[e^{-\lambda t}\right]^n = e^{-n\lambda t} \quad (3)$$

By contrast, a system consisting of parallel components works until every of its components fails. Accordingly, the reliability $R_P(t)$ of a parallel system is described by [14]:

$$R_P(t) = 1 - \left[1 - R(t)\right]^n = 1 - \left[1 - e^{-\lambda t}\right]^n \quad (4)$$

B. Duplication strategies

The primary idea of this work is to enhance the reliability of integrated systems during the design by adding redundant components. Hence, three different strategies to add redundancy were evaluated based on the design for a Wallace multiplier. Firstly, the whole multiplier is duplicated (called block duplication). Secondly, every gate of the multiplier is duplicated (gate duplication) and finally, an equal transistor was added for every transistor in the netlist (called transistor duplication). This setup guarantees a fair comparison because all three modified designs comprise the same amount of transistors. Figure 1 depicts the different design strategies and shows that the initial and the redundant components were connected to the same initial nets.

Based on the distinction between serial and parallel systems, it is assumed that the basic multiplier represents a serial configuration of components (i.e. it fails if any component fails). Furthermore, any form of introduced redundancy is treated as a parallel component. It is also to be noted that no distinction is made between n-MOSFET and p-MOSFET. Thus, following from the introduced equations (3) and (4) the basic multiplier can be described as a serial system that consists of n transistors in series:

$$R_{MULT}(t) = e^{-n\lambda t} \quad (5)$$

where λ is the failure rate of a transistor and n is the number of transistors in the netlist of the basic multiplier. Accordingly,

the reliability of the multiplier with block duplication – which is represented by two multipliers in parallel – can be derived by:

$$R_{DB}(t) = 2 \cdot e^{-n\lambda t} - e^{-2n\lambda t} \quad (6)$$

Similarly, the reliability of the multiplier with gate duplication can be calculated. Hereby, a single gate consists of a serial configuration of transistors. Such a gate is duplicated and treated as a parallel composition of the initial and the redundant gate. Finally, the doubled gates are assembled in a serial chain which leads to the following equation:

$$R_{DG}(t) = \left[2 \cdot e^{-m\lambda t} - e^{-2m\lambda t}\right]^{g} \quad (7)$$

with g being the number of gates and m being the average number of transistors per gate.

Lastly, the multiplier with transistor duplication – which represents a series of doubled transistors – is expressed by:

$$R_{DT}(t) = \left[2 \cdot e^{-\lambda t} - e^{-2\lambda t}\right]^{n} \quad (8)$$

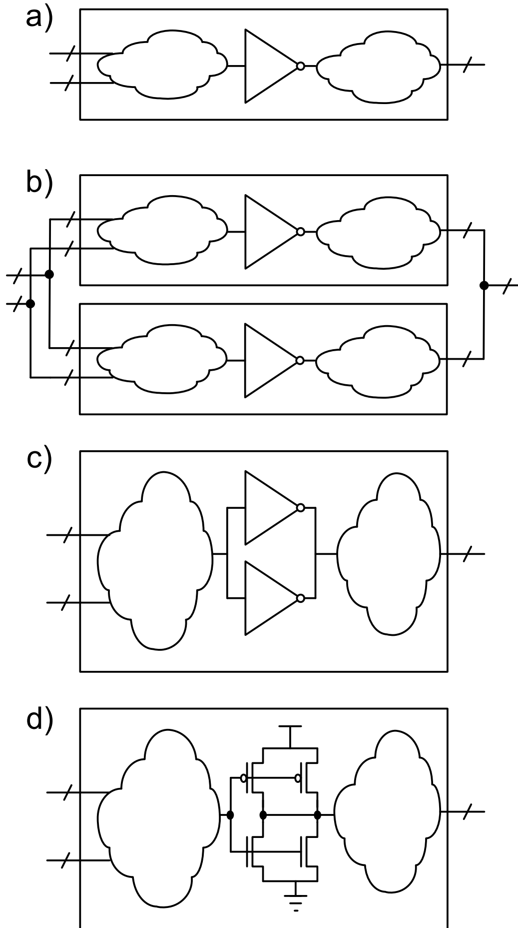

Figure 1 Duplication strategies: a) Basic multiplier b) Block duplication c) Gate duplication d) Transistor duplication

The four reliability functions based on the equations (5) to (8) – equivalently also the MTTF – are strongly dependent on the transistor failure rate λ and the number of transistors n. This means in our case for the basic multiplier that n corresponds to 408 transistors. Moreover, we defined a failure rate of $1/n$ per time unit for the simulations which conforms to $\lambda \approx 0.0025$.

C. Setup

To investigate the system reliability as regards gate oxide breakdown, a 4x4 Wallace multiplier was chosen as the reference system and was simulated on transistor level with HSpice to gain all necessary design parameters (e.g. delay, power, reliability). The reference netlist of transistors was derived from synthesis with an industrial 65 nm gate library. To be able to simulate the mechanisms of gate oxide breakdown we chose equivalent circuit models for the breakdown between gate and substrate [15]. Figure 2 represents the circuit model for oxide breakdown where the initial transistor is split into two transistors in series whereas $w = w_1 + w_2$ holds for the gate widths. The degree of the breakdown can be modeled by different values for the resistance R. A value close to zero represents a hard breakdown whereas $R >> 0$ is called a soft breakdown. Additionally, the location of the defect can be modeled by varying the ratio of w_1 and w_2. However, we assumed a hard breakdown in the middle of the gate for every defect transistor ($w_1 = w_2$). In the literature several other models exist [16][18] which claim to simulate the current flow and malfunction of the gate oxide breakdown more accurately; though, at the price of increased computational overhead and additional setup time for the simulation scenarios. Thus, for the sake of simplicity and a first attempt, we chose the first-order model of Segura et al. which needs no further investigations for the given library and which can simply be integrated into the given netlists of the multiplier.

The first set of simulations was made to compare the four designs without any defect and to evaluate the influence of the different strategies on area, delay and power consumption. After that, numerous simulations were made to investigate the designs in the presence of defects. Thereto, the defect transistors in the netlist were chosen randomly with uniform distribution and the achieved design parameters were averaged for the various simulations with the same number of defects. The achieved results will be presented and discussed in the next section.

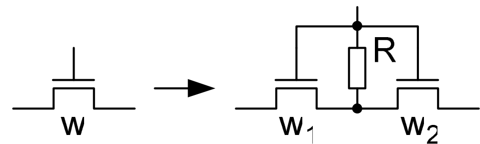

Figure 2 First-order model for an oxide breakdown between the gate and the conducting channel; w_1 and w_2 define the location; R the degree of the breakdown

IV. Results and Discussion

A. Simulations without any defect

Figure 3 depicts the results for the simulations without any defects. The delay of the redundant multipliers remains constant while the values for power and area increase by roughly factor 2 compared to the basic multiplier. This can be explained by the fact that by doubling the components the load capacitance is doubled as well. Concerning the delay, this doubling is compensated by an increased driving strength. However, power and area are directly proportional to the number of transistors and, thus, these values are twice as high as in the basic multiplier. It needs to be noted that the values for static power include classical leakage currents as well as static currents due to oxide breakdown. The impact of such currents on the overall power consumption will be explained in section IV b) and figure 7.

B. Simulations with defects

Figure 4 compares the simulated results (black lines with markers) for the reliability of the different multipliers. Furthermore, theoretical graphs for the reliability of the multipliers with transistor and gate duplication – derived from equations (7) and (8) – are depicted exemplarily to show that the simulated results correspond to the expected theoretical trend (grey lines with no markers). A similar compliance holds also true for the simulated results and the theoretical expectations of the other designs.

When examining the figure it should be considered that the time units t do not directly correspond to the number of defect transistors because equation (1) assumes a constant failure rate λ and not a constant number of failures per time unit. According to that, the number of failures per time unit decreases slightly over time. As an example, the numbers of defect transistors are plotted in the figure for two time units.

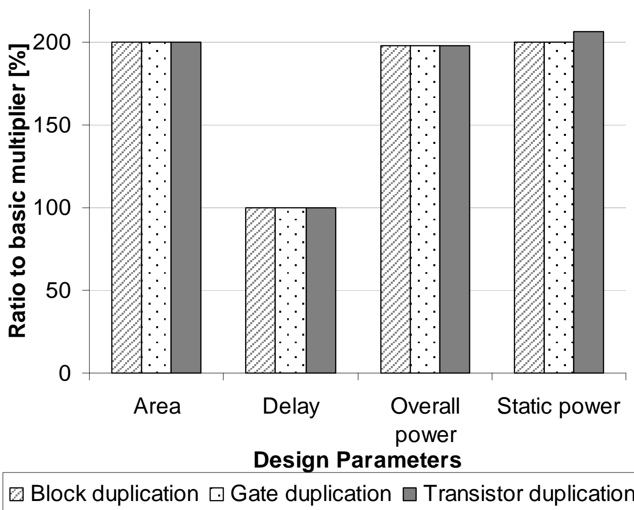

Figure 3 Results of the redundant multipliers without defects (shown is the ratio compared to the basic multiplier)

The results in figure 4 make clear that there is a significant correlation between the level of abstraction – the level where the redundant strategy is applied to – and the achieved values for the reliability. For instance, after roughly 21 time units there is still a probability of more than 60 % that the multiplier with duplicated transistors will still work, although 20 defect transistors exist in the netlist. However, for the same time instance less than 10 % of the multipliers with duplicated gates will work further on and none of the multipliers with block duplication or no duplication will work at all. To be able to simply compare the results in a quantitative manner, the MTTF was calculated based on the simulation results and equation (2). Thus, the MTTF was approximated by:

$$MTTF_{TDDB} \approx \sum_{i=1}^{a} (R(t) + R(t-1))/2 \qquad (9)$$

with a being the first time instance where all multipliers failed, and $R(t)$ being the fraction of multipliers which work correctly at time instance t.

Table 1 depicts the impressive improvements for the reliability as regards gate oxide breakdown, respectively the $MTTF_{TDDB}$. For example, the $MTTF_{TDDB}$ increases more than 25 times for the multiplier with transistor duplication compared to the basic multiplier. Furthermore, gate duplication leads to an increased expected lifetime that is after all still nearly 8–fold longer. However, block duplication exhibits even a reduced reliability which is caused by the doubled number of transistors and the fact that a single multiplier will very likely fail in the presence of one or two defects.

Table 1 $MTTF_{TDDB}$ for the duplication strategies

Duplication:	No	Block	Gate	Transistor
$MTTF_{TDDB}$:	0.537	0.446	4.237	13.552
Improvement:	0 %	-17 %	789 %	2524 %

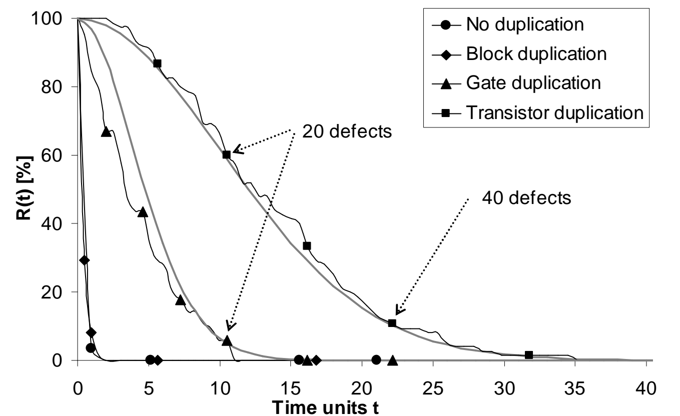

Figure 4 Comparison of simulated results for the reliability of the different multipliers (black lines) and the theoretical expectation for the trend – based on equation 7 and 8 (grey lines)

The discussed results until now did only consider the logical correctness of the different designs. However, in the presence of defect transistors a degradation of other design parameters does also have to be considered. Such results are depicted in figures 5 to 7 for the two most promising design strategies (gate and transistor duplication) where different design parameters are depicted over the number of defects. In the case of delay (see figure 5), there is a linear relation between the number of defects and the average delay of the working multipliers due to the reduced driving strength. This means that albeit the systems work functionally correct, the increase in delay might have to be compensated for to avoid timing errors. Known and approved approaches exist and are mostly applied on system level (e.g. frequency scaling).

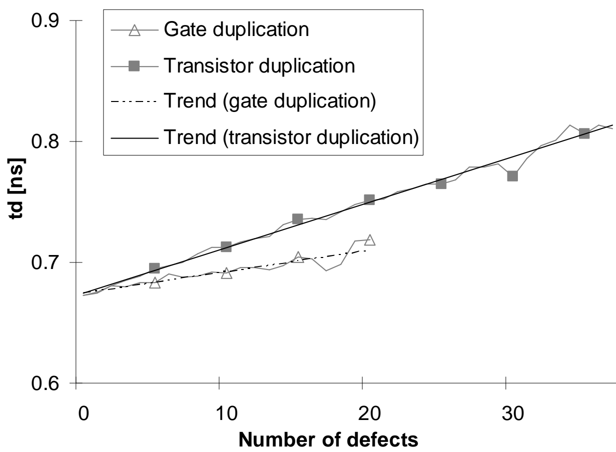

Figure 5 Average delay (t_d) for gate and transistor duplication with increasing number of overall defects

A similar characteristic can also be observed for both the overall and the static power consumption (see figure 6). The cause for the increase can primarily be found in the amount of static power caused by the defect transistors because with no defects present, static power consumption only consists of the leakage currents and is rather small. However, with larger numbers of defects the static power consumption increases because it comprises the leakage currents of all transistors plus the permanent currents through the defect gates. This issue is more clearly depicted in figure 7 which shows the rising ratio of static power to the overall power for an increasing number of defects. Moreover, it needs to be considered that – besides the improved reliability –the increase in power consumption becomes another major problem in the presence of various defects. Thus, techniques to reduce static power, like sleep transistors should also be considered [17].

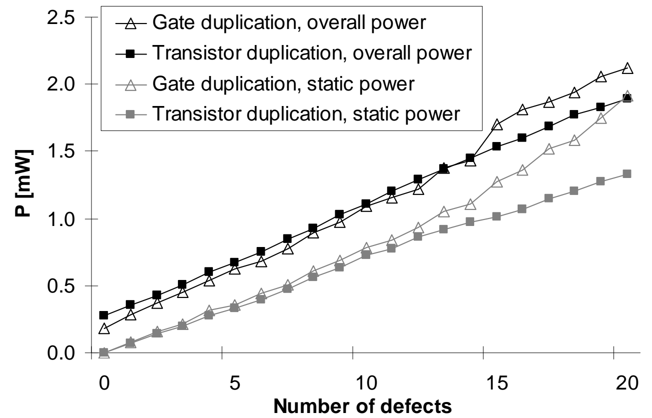

Figure 6 Power values for gate and transistor doubled multipliers

In summary, it could be shown that redundancy at the lowest abstraction level (transistor level) leads to the most reliable multipliers, followed by gate duplication. Furthermore, block duplication is not increasing the reliability. However, delay and power also gain weight with the number of defect transistors, so that system level approaches need to be considered additionally to compensate for these changes and to allow a graceful degradation of overall system performance rather than an early and sudden system failure. Hence, the gain in reliability as regards gate oxide breakdown has eventually to be traded off for other design parameters to remain with given design constraints (e.g. power, performance, area/costs). Lastly, gate duplication is to be favored over transistor duplication concerning the integration into standard design tools. Because, transistor duplication demands the creation of additional gate libraries – which is equivalent to rising costs – whereas gate duplication just requires changes in the compiler strategies.

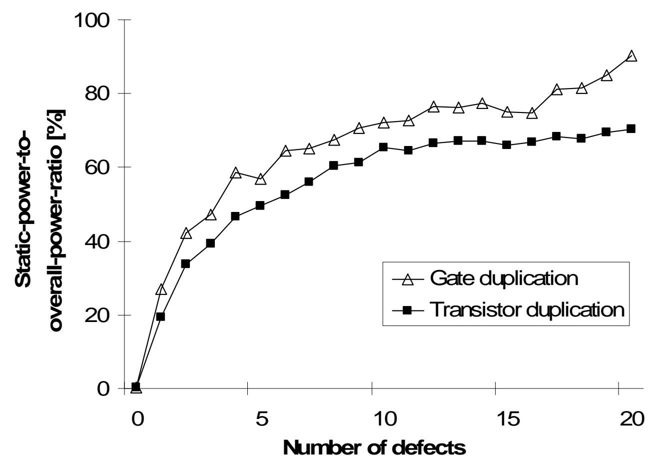

Figure 7 Ratio of static to overall power for an increasing number of defect transistors

V. Conclusion

This contribution identified the need for improvements of lifetime reliability. Therefore, three different approaches which add redundancy to enhance reliability against gate oxide breakdown at different design levels were introduced and investigated thoroughly. The conducted simulations with circuit models for hard gate oxide breakdown demonstrated that the lowest abstraction level (here transistor level) promises the most impressive improvements for system reliability. Such improvements were achieved at constant delay but at the price of increased area and power consumption. Moreover, it was pointed out that in the presence of defects delay and power consumption degrade which needs to be compensated by system level approaches to avoid the increase of other failure mechanisms (e.g. timing errors, overheating). Finally, gate duplication should be favored due to the good results for reliability and the possibility for simple integration into existing design flows and CAD tools.

Future efforts aim at implementing devices with different gate oxide thicknesses as well as more elaborate and complex breakdown models to additionally evaluate the impact of soft breakdowns.

References

[1] Srinivasan, J., Adve, S., Bose, P. and Rivers, J., "The Impact of Technology Scaling on Lifetime Reliability", In Proc. of DSN, 2004.

[2] Stathis, J., "Reliability Limits for the Gate Insulator in CMOS Technology", In IBM Journal of Research and Development, 2002.

[3] Crook, D., "Method of Determining Reliability Screens for Time Dependent Reliability Breakdown", In Proc. of Intern. Reliability Physics Symposium, 1979.

[4] Vogel, E. et al., "Reliability of Ultra-Thin Silicon Dioxide Under Combined Substrate Hot Electron and Constant Voltage Tunneling Stress", In Trans. of Electron Devices, vol. 47, no. 6, 2000.

[5] Semiconductor Industry Association (SIA), "International Technology Roadmap for Semiconductors", Release 2007, Published on-line: http://www.itrs.net/.

[6] Chen, Z. and Koren, I., "Techniques for Yield Enhancement of VLSI Adders", In Proc. of ASAP, 1995.

[7] Chiluvuri, V. and Koren, I. "Layout-Synthesis Techniques for Yield Enhancement", In Trans. of Semic. Manufacturing, vol. 8, no. 2, 1995.

[8] Omana, M., Rossi, D. and Metra, C., "Latch Susceptibility to Transient Faults and New Hardening Approach", In Trans. on Computers, vol. 56, no. 9, 2007.

[9] Mitra, S. et al., "Robust System Design with Built-In Soft-Error Resilience", In Computer, vol. 38, 2005.

[10] Srinivasan, J. et al., "The Case for Lifetime Reliability-Aware Microprocessors", In Proc. of ISCA, 2004.

[11] Sirisantana, M., Paul, B. and Roy, K., "Enhancing Yield at the End of the Technology Roadmap", In Trans. of Design&Test of Computers, vol. 21, no. 6, 2004.

[12] Cornelius, C., Sill, F., Saemrow, H., Salzmann, J., Timmermann, D., da Silva, D., „Encountering Gate Oxide Breakdown with Shadow Transistors to Increase Reliability" , In Proc. of SBCCI, 2008.

[13] Koren, I. and Krisha, C., "Fault-tolerant Systems", M Kaufmann, 2007.

[14] Romeu, j., "Understanding Series and Parallel Systems Reliability", In START, Vol. 11, nr. 5, 2005.

[15] Segura, J., Benito, C., Rubio, A. and Hawkins, C., "A Detailed Analysis of GOS Defects in MOS Transistors: Testing Implications at Circuit Level", In Proc. of Test Conference on Driving Down the Cost of Test, 1995.

[16] Renovell, M., Gallière, J., Azaïs, F. and Bertrand, Y., "Modeling the Random Parameters Effects in a Non-Split Model of Gate Oxide Short", In Journal Electronic Testing, vol. 19, no. 4, 2003.

[17] Johnson, M. C., Somasekhar, D., Roy, K., "Leakage control with efficient use of transistor stacks in single threshold CMOS", In Proc. of DAC, 1999.

[18] Kaczer, B., et al., "Impact of MOSFET Gate Oxide Breakdown on Digital Circuit Operation and Reliability", In Transactions on Electron Devices, vol. 49, no. 3, 2002