Twin Logic Gates – Improved Logic Reliability by Redundancy concerning Gate Oxide Breakdown

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

Department of Electrical Engineering, University of Rostock, Rostock, Germany

*Department of Electrical Engineering, Federal University of Minas Gerais, Belo Horizonte, Brazil

Email: hagen.saemrow@uni-rostock.de or franksill@umfg.br

ABSTRACT

Because of the aggressive scaling of integrated circuits and the given limits of atomic scales, circuit designers have to become more and more aware of the arising reliability and yield concerns. So far, only very little research efforts have been put into low-level approaches for lifetime reliability, whereas lots of efforts have focused on soft-errors and system-level solutions. In this paper, we introduce and compare three diverse design approaches which apply redundancy on different abstraction levels to enhance the reliability of a Wallace multiplier as regards gate oxide breakdown. The results of the test design were further improved by adding transistors and gates with different gate oxide thicknesses. The achieved results show that lifetime reliability increases up to 200 % at constant delay by adding redundant gates, subsequently called Twin Logic Gates. However, this comes at the price of overhead for area as well as power consumption. Furthermore, it needs to be noted that the presented strategies can additionally improve defect yield.

Categories and Subject Descriptors

B.5.3/B.7.3 [**Register-Transfer-Level Implementation**] / [**Integrated Circuits**]: Reliability and Testing – *redundant design*

Keywords

Integrated circuit design, redundant systems, reliability, gate oxide breakdown.

1. MOTIVATION

Enhancing yield has always been a major concern in the manufacturing process of integrated circuits. However, manufacturing engineers as well as designers are recently facing more and more severe lifetime reliability issues. By now, aggressive scaling of integrated circuits has led to increased sensitivity of transistors and interconnects to different kinds of failure mechanisms during system operation and manufacturing as we are approaching the limits of nanotechnology. Besides decreased design and process error margins due to defects and material imperfections (which cuts down on yield), non-ideal scaling, increased transistor count and alarming power density levels also result in a decline of lifetime reliability due to failure mechanisms such as Time-Dependent Dielectric Breakdown (TDDB), electromigration or thermal cycling [1]. Although known for a long time, researchers and engineers have not yet investigated all relevant facts about these undesired effects. Due to the continuous scaling, gate oxide — the dielectric isolation between the transistor input and the conducting channel — has become highly vulnerable to breakdown mechanisms causing transistor defects and logical system malfunctions. A gate oxide breakdown (GOB) is defined as the process when a conducting path emerges between the gate and the substrate or source/drain, respectively [2]. This could result from two different effects. Firstly, extreme overvoltage, e.g. caused by Electro-Static Discharge, leads to a sudden damage. Secondly, a rather slow destruction over time is also possible, called TDDB. Thereby, an autocatalytic loop of events takes place. Overlapping charge traps create a conducting path between gate and substrate (or source/drain) which leads to increased current flow and heat dissipation. This 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]. Depending on the I-V characteristics of the defect transistor, the breakdown is distinguished into soft breakdowns and their final result: resistor-like hard breakdowns. The conductance of the soft breakdown is strongly non-linear and more limited concerning the amount of current flow through the gate [5].

Besides the identification of the soft breakdown mode, it has been shown in the mid 1990's that gate oxide breakdowns not necessarily result in logic malfunctions of the transistor as assumed before. Rather than a logic failure, an affected transistor and its associated logic cell suffer from a modified delay [5]. Certainly, the whole circuit fails if the timing between the cells is no longer balanced due to delay failures of multiple gate oxide breakdowns.

Because of the rising transistor count per die, the number of system failures caused by gate oxide breakdown will increase with every new technology. Unfortunately, full functional system tests are not feasible due to the rising complexity of integrated systems. Thus, tool assisted insertion of reliability mechanisms into the design flow will be one of the key priorities in the future [6]. In this contribution, different strategies for inserting redundancy are presented which can more or less easily be embedded into current CAD tools.

Low-level redundancy techniques which enhance defect yield are already established in integrated 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 [7]. By contrast, little effort has been made so far on techniques to enhance lifetime

SBCCI'09, August 31st - September 3rd, 2009, Natal, RN, Brazil.
Copyright 2009 ACM 978-1-60558-705-9 $5.00.

reliability. One such high-level approach was published in [8] 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 was made in [9] 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 [10] where the redundant Shadow Transistors were inserted only at those instances that are most vulnerable to TDDB which increases not just the yield but also lifetime reliability. By inserting transistors with higher gate oxide thickness than the original ones, reliability could be increased even more.

Based on the promising results for redundant transistors, this paper compares the results of redundancy enhancements for a Wallace multiplier and different levels of abstraction. Higher abstraction levels than the transistor level have been chosen 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 various levels of abstraction by presenting figures on reliability, delay and power consumption. Additionally, further results will be shown for the insertion of transistors with higher gate oxide thickness.

2. APPROACH

2.1 Fundamentals of Reliability

There are three important parameters for the analysis of reliability. Firstly, the failure rate λ represents the rate at which an individual component suffers from faults. Hence, the failure rate of a system is based on the failure rates of its individual components [11]. In this contribution the multiplier represents the system and the transistors its components with an equal failure rate for all transistors, except for transistors with different gate oxide thicknesses as described in chapter 2.2. Besides the dependence on technology and further circumstances like ambient temperature or gate voltage level, the failure rate of the components — and thus, also of the system — depends on the current age. The relation to the age is captured in figure 1 which is well known as the bathtub curve. There are three different time intervals. The first one is called infant mortality phase in which systems with manufacturing defects, sensitive or borderline components tend to fail. Because of this, the failure rate at the beginning of a system's lifetime is very high. This phase (1) directly affects yield and is the reason for burn-in tests at the end of manufacturing. As time goes on, the failure rate decreases with the rejection of these weak systems and the circuits work mostly correctly as the failure rate is very low and roughly constant in this phase (2). Lastly, systems reach the wear-out phase (3) where aging effects like TDDB raise the failure rate again [11].

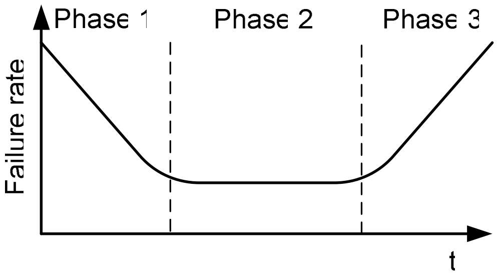

Figure 1 Reliability bathtub curve

The reliability *R(t)* of a component itself is the probability of the device to perform as desired until time *t*. Although the assumption of a constant failure rate is sufficient for many reliability calculations, figure 1 shows that this not true in every case. In such cases, the failure rate is time-dependent and will be calculated with the Weibull distribution. The corresponding reliability can be calculated by [11]:

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

where β is an accelerating or decelerating factor. The failure rate decreases (phase 1) for $\beta < 1$, remains constant (phase 2) for $\beta = 1$ and increases (phase 3) for $\beta > 1$. The system's reliability is a result of the reliability of its components which affect the reliability of the system due to for instance delay shifts, logical failures or increased power consumption.

Closely related to the probabilistic term for reliability is the Mean Time to Failure (MTTF) which is the average time a system operates until it fails. It is equal to the expected lifetime if the system cannot be repaired — as expected in most cases for integrated circuits. It can be calculated by [11]:

$$MTTF = \int_0^{\infty} R(t)dt \qquad (2)$$

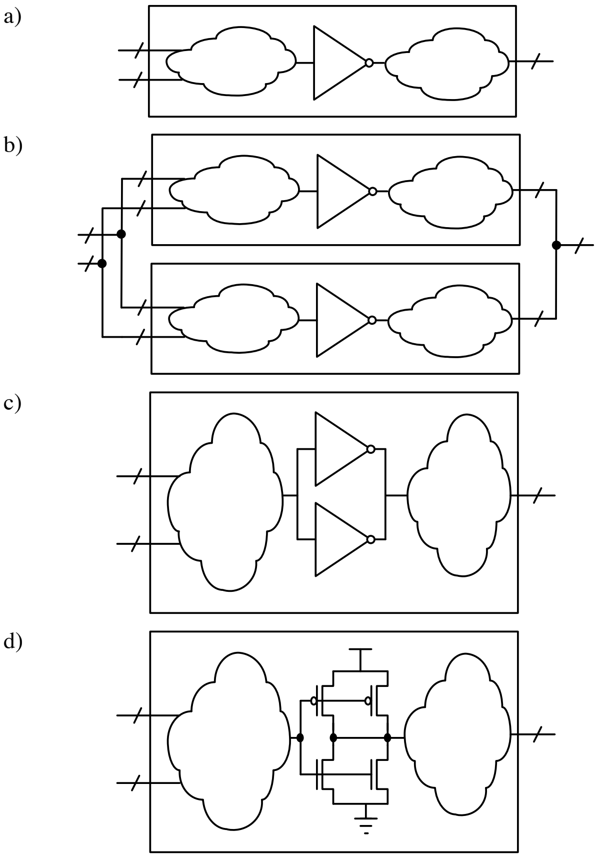

Figure 2 Duplication strategies: a) Basic Multiplier b) Block duplication c) Gate duplication (Twin Logic Gates) d) Transistor duplication

2.2 Strategies for Reliability Improvements

Enhancing the reliability of a system by adding redundant components is the key idea of this work. Therefore, three different redundancy strategies were evaluated based on the design for a Wallace multiplier. Firstly, the whole multiplier was duplicated (called block duplication). Secondly, every gate of the multiplier was duplicated (Twin Logic Gates) 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 2 depicts the different design strategies and shows that the initial and the redundant components were connected to the same initial nets.

Another enhancement is the use of transistors with thicker gate oxides because gate oxide breakdown occurs more likely if the gate oxide is thinner. Therefore, every duplication strategy was implemented twice. One configuration consists only of transistors with equal oxide thicknesses whereas the duplicated transistors of the second implementation have thicker gate oxides. Except for the gate oxide thickness, all other design parameters (e.g. gate width) remained identical as in the basic multiplier. As a rule of thumb, it is assumed that with a 0.1 nm thicker gate oxide the failure rate decreases by one magnitude which is a widely accepted presumption [2]. For our implementations this rule of thumb results in a 1.2 times higher failure rate for transistors with standard oxide thickness (SVT – Standard Threshold Voltage transistors) than for transistors with thicker gate oxide (HVT – High Threshold Voltage transistors).

2.3 Simulation Setup

A 4x4 Wallace multiplier was chosen as the reference system. The essential design parameters, like delay, power and reliability, were extracted on transistor level with HSpice. The netlist of transistors was derived from synthesis with an industrial 65 nm gate library. As mentioned in chapter 2.2 two transistor types – HVT and SVT – were used for the gate synthesis. To be able to simulate the mechanisms of gate oxide breakdown, we chose equivalent circuit models which is depicted in figure 3 [13] where two transistors were added in parallel. Modifying the widths w_1, w_2 and w_3 of all three transistors will result in an appropriate representation of the current flow on drain, source and gate for a transistor with a punctual gate oxide breakdown on a certain location at the transistor gate. The resistor R connecting the drains of the two additional transistors with the gates of all three transistors simply represents the resistance of the short between gate and bulk of the transistor. With this model it is possible to simulate a gate oxide breakdown on any horizontal and vertical location between gate and substrate with any resistance of the breakdown path [13].

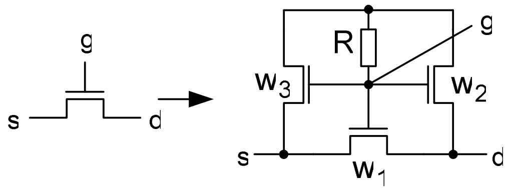

Figure 3 Cicruit model for a gate oxide breakdown between the gate (g) and the conducting channel; s- source; d- drain

The first set of simulations compared the behavior of all implemented designs in the absence of any defect to evaluate the influence of the different strategies on area, delay and power consumption. After that, numerous simulations were made to investigate the designs with defects. Thereto, the defect transistors in the netlists were chosen randomly with uniform distribution, considering the difference between HVT and SVT. Further on, the defect location on the gate oxide was also chosen randomly. The achieved design parameters were averaged for the various Monte-Carlo simulations with the same number of defects. Two different types of defect simulations were made. Firstly, simulations with a constant failure rate represent the designs in phase 2 of figure 1. There, hard breakdowns (R = 1 Ω) were being inserted one by one during these investigations until a system failure was detected. The second type of simulations represented the behavior of the design with an increased failure rate as found in phase 3 of figure 1. Therefore, every tenth transistor in the duplicated designs was chosen to become defect. To behave like non-defect transistors at the beginning of these simulations, every defect had at the start a very high resistance R. These resistances were decreased over time to simulate a soft breakdown developing into a hard breakdown after a certain time interval. The modifying parameters were calculated randomly but also in such a way that an exponential growth rate of the current flow of the defect transistor per time was given.

3. RESULTS & DISCUSSION

3.1 Strategy comparison without defects

Figure 4 depicts the results for the simulations without any defects. As expected, the delay of the redundant multipliers with SVT transistors remains constant, whereas the redundant multipliers with HVT transistors are slower than the basic multiplier. The reason is that the doubled load capacitance of the redundant multipliers can be compensated by the increased driver strength for the SVT strategy; meanwhile the weaker HVT components are not able to counterbalance this increase completely. The larger load capacitance is also the reason for the roughly doubled overall power consumption for the redundant multipliers. Without the existence of any defect the overall power is dominated by the dynamic power consumption which is linearly dependent on load capacitance.

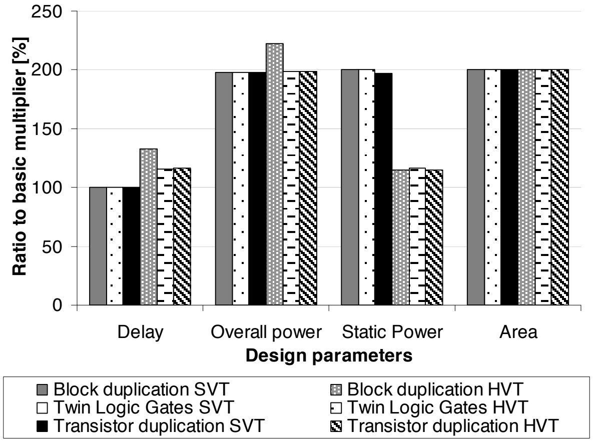

Figure 4 Results of the redundant multipliers without defects (shown is the ratio compared to the basic multiplier)

Furthermore, solely the block duplication (with HVT) exhibits slightly higher values for the delay and power dissipation as the other redundant multipliers. The reason is that the more components are included inside a redundant block - which means the larger the difference is in performance and power due to the SVT and HVT devices - the more is the node affected where both blocks are connected together again. Thus, the differences of the timing and output slopes of the two blocks and, hence, the partial contrary switching of the output of the whole multiplier leads to an increased delay and also to an elevated influence of short circuit currents which raise the overall power consumption. Another design parameter which is affected by the redundancy strategies is the static power consumption which comprises classical leakage currents and static currents due to oxide breakdown as well. Without the presence of defects the static power consumption is also increased by roughly factor 2 for the SVT redundant multipliers because of the doubled number of transistors. By contrast, the static power consumption of the HVT redundant multipliers is only raised by about 15 % which is due to the higher threshold voltage of HVT transistors that results in lower leakage currents. As a last parameter the area is doubled for every redundant multiplier because of the doubled number of transistors.

3.2 Defect Simulations

The results of functional reliability analyses are depicted in figure 5. Due to the constant failure rate λ (phase 2) and not a constant number of failures per time unit, it should be considered when examining the figure that the number of failures per time unit decreases slightly. As an example, after 50 time units 47 defects were inserted into the SVT-doubled designs, whereas one defect appeared after one time unit. The frequency of the DUT was set to the double time interval of the critical path of the basic multiplier, which means that a defect occurs also if the critical path of the DUT exceeds this timing constraint. Unlike expectations due to former simulations with older circuit models [15] and due to the assumption that the more fine grained the redundancy is the more reliable the design is, the most reliable designs are the multipliers with duplicated gates. As depicted in figure 5, multipliers with Twin Logic Gates provide a very good enhancement for the lifetime reliability of the multiplier. For instance, after roughly 50 time units there is still a probability of more than 81 % that the multiplier with duplicated HVT gates (77 % for SVT respectively) will still work, whereas only 2 % of the basic multipliers will still function correctly. In addition, at this time instance 59 % of the multipliers with HVT transistor duplication still work (33 % for SVT transistor duplication). Further on, multipliers with block duplication provide only little reliability enhancement of 7 % (HVT) or even decrease reliability in proportion to the basic multiplier (every design with SVT block duplication is defect at this point in time). Generally, it can be considered that multipliers with Twin Logic Gates are the best choice for reliability enhancement with a completely doubled design. Furthermore, designs with redundant HVT components provide higher reliability than designs with doubled SVT components which is due to the lower defect probability of HVT transistors.

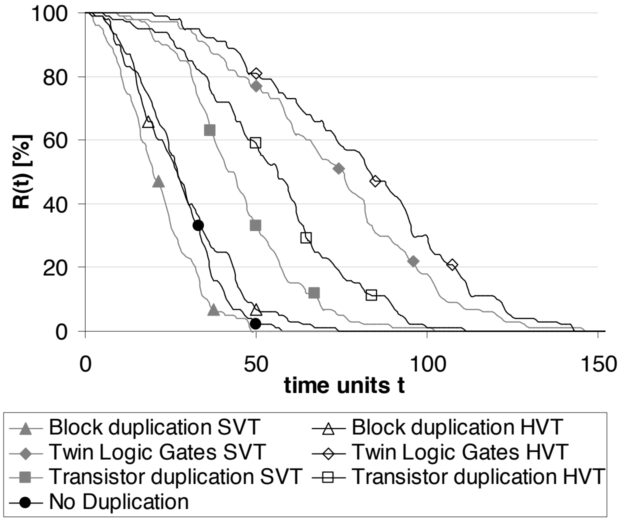

Figure 5 Comparison of simulated results for the reliability of different multipliers (phase 2)

Furthermore, it is remarkable that the basic multiplier has an intrinsic reliability against GOB which leads to a design which not automatically fails once a GOB occurs. Therefore, the drawbacks of the duplication strategies like increased area and increased power consumption should be considered against the fact that some reliability is given by the design itself.

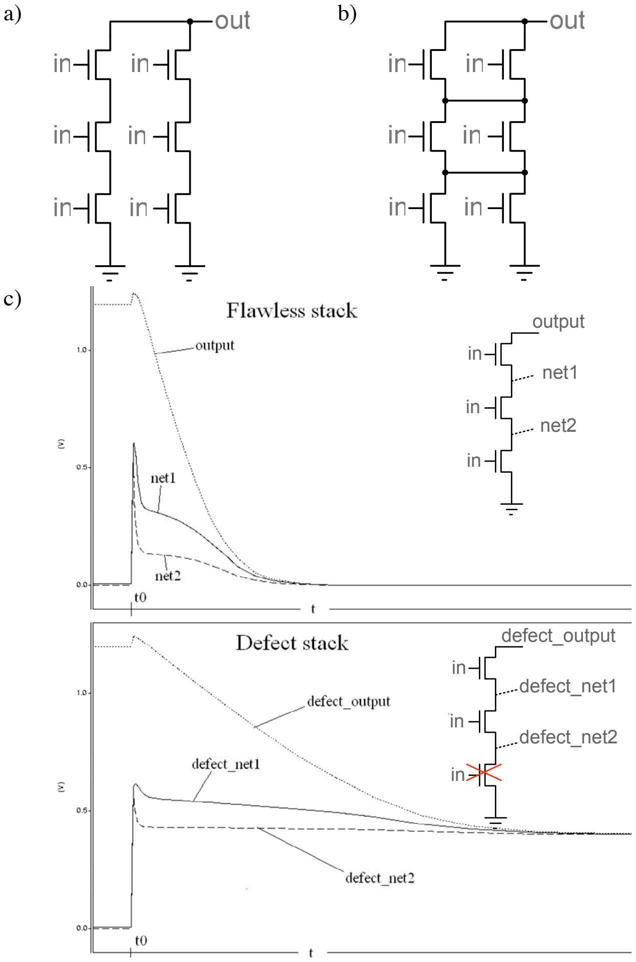

Figure 6 NMOSFET stacks: a) Duplicated stack (Twin Logic Gates) b) Duplicated stack (Transistor Duplication c) Voltage levels at the nets of the flawless and defect NMOSFET stack

The reason for the better results for Twin Logic Gates in contrast to transistor duplication are found in the transistor stacks due to the fact that both implementations only differ in the duplication of transistor stacks as can be seen exemplarily in figure 6 where a stack of 3 N-MOSFET was duplicated according to a) gate duplication and b) transistor duplication. Figure 6 illustrates the differences between a flawless stack and a stack with a GOB defect at the bottom transistor on the basis of voltages at the nodes between the transistors (c). Before time t0 the output net is charged high and all other nodes are at low level. At time t0 the inputs rise to high which leads to a discharge of the output nodes. First, the voltages of the nodes net1 and defect_net1 are rising. Due to increasing voltages at these nodes and the resultant voltage difference between drain and source of the middle transistor, net2 and defect_net2 are also charged. This leads to a current flow through the bottom transistor which finally results in a connection to ground. As a consequence, net2 and defect_net2 are discharged, similarly net1, defect_net1 and the outputs. Both cases differ in the velocity with which both outputs can be discharged (fall time). Due to the defect at the bottom transistor, defect_net2 is charged to a voltage which is related to the degree of the GOB at this transistor when the inputs switch to high level. As a result, the current flow from drain to source at the middle transistor is rather pinched off which finally leads to an increased fall time for the defect stack and an incomplete discharge of the output. Thus, if a defect occurs in a doubled stack, one of both Twin Logic Gate stacks will remain unaffected and is able to discharge the output faster and almost completely. By contrast, transistor duplication leads to a slower and more incomplete discharge of the output due to the cross links to the parallel stack. Therefore, stacks duplicated with Twin Logic Gates are slightly faster than transistor doubled stacks.

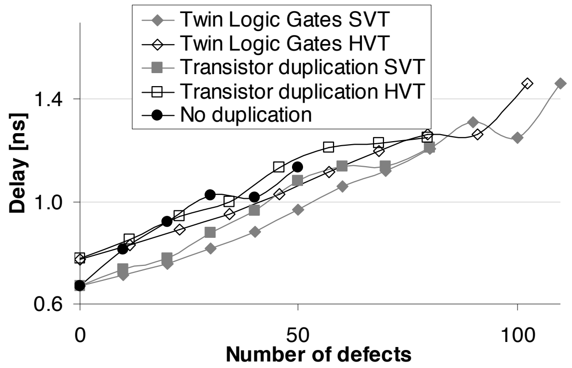

Figure 7 Average Delay for the basic multiplier and the multipliers with gate and transistor duplication

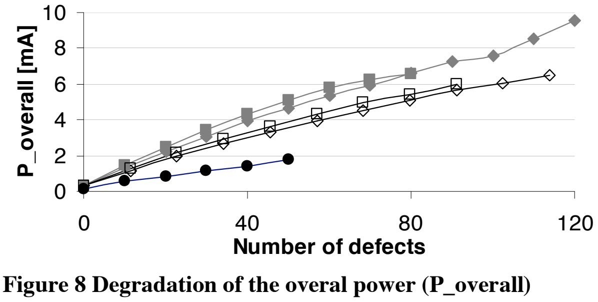

Figure 8 Degradation of the overal power (P_overall)

Lastly, this leads to better reliability results for Twin Logic Gate designs because they meet the voltage level and timing constraints longer in the presence of multiple defects. To compare the strategy improvements in a quantitative manner, the MTTF of every design was calculated based on the simulation results and equation (2). Thus, the MTTF was approximated by:

$$MTTF \approx \sum_{i=1}^{a} (R(t) + R(t \quad 1))/2 \qquad (3)$$

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

Table 1 MTTF for the duplication strategies

Duplication	Transistor type	MTTF	Improvement
No	-	26.72	0 %
Block	SVT	21.17	-21 %
	HVT	28.42	6 %
Transistor	SVT	44.34	66 %
	HVT	54.71	105 %
Gate	SVT	72.52	171 %
	HVT	80.97	203 %

Table 1 depicts the reliability enhancements for the MTTF. The expected lifetime of the Twin Logic Gates design (HVT) is three times longer than the lifetime of the basic multiplier whereas transistor duplication (HVT) increases the reliability only up to more than 100 % concerning the MTTF. Furthermore, this table clearly depicts the marginal improvement of the HVT block duplication (~6 %) and even the reliability decrease with SVT block duplication. This fact is due to delay differences between both blocks caused by variable defect numbers and cell locations in both blocks which lead to undefined output signals for the whole multiplier.

Until now only the logical correctness was discussed. However, in the presence of defect transistors the other design parameters will be degraded as well. However, the results show a graceful degradation behavior for all designs. They are depicted in the figures 7 to 8 for the most promising design strategies. Figure 7 shows the development of the delay for an increasing numbers of defects.

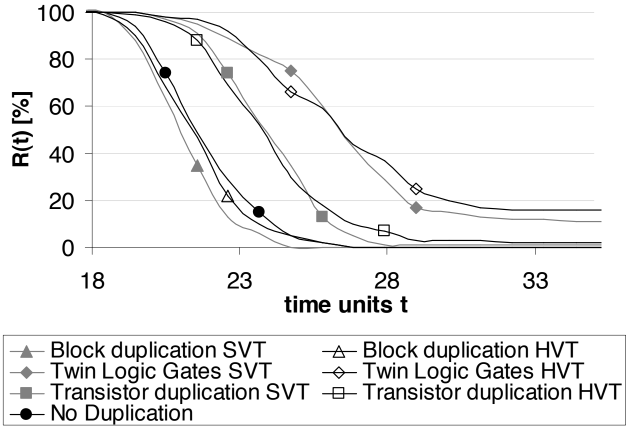

Figure 9 Comparison of simulated results for the reliability of the different multipliers (wear-out phase)

There is an approximate linear relation between the number of defects and the average delay of the working multipliers caused by the reduced driving strength of the suffering transistors. Due to the slower HVT transistors these doubled designs are slower as regards the absolute values. Furthermore, the basic multiplier with no duplication at all will degrade faster in performance due to no redundant components which could partially compensate for the system's performance loss caused by defect transistors. Figure 8 depicts a similar relation between the increased overall power consumption and the rising number of defects. Besides the basic multiplier that consumes roughly half of the power compared to the duplicated designs, the HVT designs exhibit lower overall power consumption. The reason for the increased power dissipation of all designs is primarily due to the increase of static power consumption caused by defect transistors. With no defects present, the static power consists only of leakage currents which are rather small in contrast to the dynamic power in our case. However, with larger numbers of defects the static power consumption increases, comprising now leakage currents and the permanent current flow through the defect transistor gates. The impact of these currents on the overall power consumption is dominating the overall power consumption as soon as the number of defect transistors increases. Therefore, redundant designs with HVT transistors consume less overall power because of their lower current flow due to their thicker gate oxides. Nevertheless, it needs to be considered that this increase in power consumption represents a major problem in the presence of various defects. Thus, techniques to reduce static power like sleep transistors should be considered additionally [13].

Similar reliability results and a graceful degradation behavior were achieved for the wear-out simulations. As depicted in Figure 9 the characteristic of the reliability curves of every duplication strategy remains the same as for the phase-2 simulations. But due to the simultaneous defects of a lot of transistors and therefore the fast malfunction of most multipliers, the differences are not that huge as for the investigations before. The MTTF improvements, which are strongly dependent on the number of chosen transistors, range from -3 % (SVT block duplication) to more than 30 % (HVT Twin Logic Gates). It is also observable that some multipliers still work after a GOB at all chosen transistors.

In summary, the simulations show a significant improvement of reliability for phase 2 of a system's lifetime if the gates are doubled, especially Twin Logic Gates featured with HVT transistors. In the wear-out phase reliability enhancements were still available with the favored Twin Logic Gates strategy, but not to the extent of phase 2. However, the improvements offer new possibilities for automatic CAD tools because of the simple integration of gate based strategies into the design flow, in contrast to transistor duplication strategies. The improvements were achieved at constant or slightly raised delay but at the price of increased area and power consumption. Though, the improvements can also raise defect yield because of the redundancy itself. It is remarkable that the original design with no duplication strategy shows a basic reliability which could compensate some defects. However, the further improvements for yield and the slower degradation of the duplication strategies cannot be achieved accordingly.

Summing up, both gate and transistor duplication strategies are able to lower the system failure rate curve of figure 1 because of the following reasons. Firstly, the strategies gain a better yield in phase 1. Secondly, the reliability enhancements lead to a lower system failure rate and an extension of phase 2 and, finally, wear-outs in the last phase will be delayed for a certain time.

4. CONCLUSION

This contribution identifies the needs for improvements of lifetime reliability. Therefore, six different approaches which add redundancy to enhance reliability against gate oxide breakdown at different design levels and featured with transistors with different gate oxide thicknesses were introduced and investigated thoroughly. The simulations with different failure rates and corresponding circuit models for hard gate oxide breakdown demonstrated that a design with duplicated gates (Twin Logic Gates) promises the most impressive enhancements of system reliability. This provides the possibility of a simple integration into existing design flows and CAD tools. The improvements were further elevated by the usage of transistors with thicker gate oxides for the duplication. All enhancements 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 failures. Future efforts aim at identifying critical gates and implementing an approach which only doubles these gates to reduce the area and power consumption overhead and the use of benchmark circuits.

5. REFERENCES

[1] Srinivasan, J., et al., "The Impact of Technology Scaling on Lifetime Reliability", DSN, 2004.

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

[3] Crook, D., "Method of Determining Reliability Screens for Time Dependent Reliability Breakdown", IRPS, 1979.

[4] Vogel, E. et al., "Reliability of Ultra-Thin Silicon Dioxide Under Combined Substrate Hot Electron and Constant Voltage Tunneling Stress", Electron Devices, 2000.

[5] Kaczer, B. et al., "GOB in FET devices and circuits: From nanoscale physics to system-level reliability", Elsevier, 2007.

[6] Semiconductor Industry Association (SIA), "International Technology Roadmap for Semiconductors", Release 2007,

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

[8] Srinivasan, J. et al., "The Case for Lifetime Reliability-Aware Microprocessors", ISCA, 2004.

[9] Sirisantana, M., et al., "Enhancing Yield at the End of the Technology Roadmap", Design&Test of Computers, 2004.

[10] Cornelius, C., et al., "Encountering GOB with Shadow Transistors to Increase Reliability", SBCCI, 2008.

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

[12] Lindner, B. et al., "Growth and Scaling of Oxide Conduction after Breakdwon", IRPS, 2003.

[13] Renovell, M., et al., "Modeling the Random Parameters Effects in a Non-Split Model of GOS", Electronic Testing '03.

[14] Johnson, M. C., et al., "Leakage control with efficient use of transistor stacks in single threshold CMOS", DAC, 1999

[15] Saemrow, H., et al., "Comparison of Strategies for Redundancy to improve Reliability concerning GOB", TUZ, 2009