Selective redundancy to improve reliability and to slow down delay degradation due to gate oxide breakdown

Hagen Saemrow, Claas Cornelius, Philipp Gorski, Andreas Tockhorn and Dirk Timmermann
Department of Electrical Engineering, University of Rostock, Rostock, Germany
Email: hagen.saemrow@uni-rostock.de

Abstract— **Because of the aggressive scaling into the nanometer regime, degradation due to wearout significantly impairs design parameters. For instance, such wearout is caused by gate oxide breakdown, which decreases the operating lifetime of integrated circuits to an extent that cannot be neglected by circuit designers to date. In this paper, we introduce an approach which applies selective redundancy to different combinational designs in order to improve reliability as regards gate oxide breakdown. Therefore, the most vulnerable transistor stacks of standard cells are doubled based on activity and the propagation delay of the design. Finally, reliability improvements of up to 75 % are presented that are gained with Spice simulations. Such improvements come at the price of overhead for area and power consumption as well as delay of at most 14 %. However, it is interesting to notice that the initial delay penalty of our enhanced designs finally turn into a timing advantage, as the designs are more and more affected by wearout over time. Hence, this advantage translates into further reliability improvements when clock requirements are also considered. Besides, it needs to be noted that the presented strategies can additionally improve defect yield.**

Keywords: Integrated circuits, redundant systems, reliability, gate oxide breakdown, delay degradation.

I. MOTIVATION

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 [1]. Depending on the changing I-V characteristic of a defect transistor, the gate oxide breakdown (GOB) is distinguished into the initial soft breakdown and the final result: the resistor-like hard breakdown. The shift of the conductance during soft breakdown is strongly non-linear and the amount of current flow through the gate is smaller compared to the final hard breakdown [2]. It has been shown that GOBs not necessarily result in logic malfunctions of the transistors. More than resulting in logic failures, an affected transistor and its associated logic cell suffer from modified delay [3]. Certainly, the whole circuit fails if the timing between the cells is no longer balanced due to delay failures of multiple GOBs or when the extent of GOBs changes voltage levels due to conduction to the power lines V_{DD} or ground V_{SS}.

Unfortunately, full functional system tests are not feasible to face the rising reliability issues due to the immense and further 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 [4]. The acknowledged assumption that device failures cannot be prevented but have to be resolved has also been adopted for this contribution. Based on this motivation, [5] presented an approach where redundant transistors are randomly inserted into the design to increase yield as regards stuck-open transistors. This idea has been extended in [6] where redundant Shadow Transistors are inserted only at those instances that are most vulnerable to TDDB. This approach not just increases the yield but also lifetime reliability. As such works rest upon transistor level, redundancy enhancements on different levels of abstraction have been compared in [7] in order to allow the transfer from such transistor-level approaches to common CAD tools and given gate libraries. It has been shown that approaches on gate level just as well provide large improvements of reliability compared to non-redundant designs. Against this background, this contribution shows that significant reliability enhancements can be achieved when redundant transistor stacks are selectively inserted, whereas such stacks can easily be transferred to gate level libraries. Furthermore, thorough investigations based on Spice simulations also show a decrease of the worst-case delay in the presence of GOBs. This result can be very useful in order to stabilize clock-limited paths during wearout. Hence, a longer operating lifetime of the circuit can be achieved.

II. PARTIALLY REDUNDANT TRANSISTOR STACKS

The fundamental idea of redundant transistors (which are connected in parallel to their counterpart) is that penalties due to GOB in one of both transistors can be alleviated by the second one. The intact transistor can thus counteract the increased transition and propagation delay as well as the worsened voltage level at the drain net. In order to lower the overhead due to redundancy in terms of area and power consumption, we propose to only insert redundancy selectively to enhance both reliability and performance of a design in the presence of GOB. Accordingly, only those transistors whose inputs are most susceptible to GOB are doubled. As the operating lifetime *ttf* with respect to oxide breakdown is dependent on the existence of an electric field affecting the gate (which is equivalent to the voltage level at the gate input), the *ttf* can be formulated as:

$$ttf \propto {V_{GS}}^{a+bT} \tag{1}$$

where V_{GS} is the gate-source voltage, T the temperature and a and b are technology dependent parameters [8]. The gate-source voltage again leads to a dependency of the operating lifetime on the probability P_{ON}, which reflects the voltage level that turns the transistor on and which are easily extracted from activity files of RTL-simulations:

$$ttf_{nMOS/pMOS} \propto P_{ON} = \frac{t(V_{GS} \approx V_{DD}/V_{SS})}{t_{OPERATING\_TIME}} \tag{2}$$

One drawback of simple redundancy is that transistors and their doubled counterpart are influenced by the same charge imposition over time. Hence, our proposal focuses on

additional structural impacts in order to electrically disconnect a redundant transistor from the switching net (called standby phase). By contrast, the redundant transistor is reconnected to the switching net when the performance and reliability of the design need to be improved (called enhance phase). The switching from the standby to the enhance phase is managed within higher system levels when these instances notice significant changes as regards reliability or performance of a sub-circuit. Appropriate indications for the higher level management can be incorrect or delayed results. How and when the switching is performed is not part of this contribution.

It is shown in [7] that standard cells benefit more from redundant (but separated) transistor stacks than from redundant individual transistors because the current paths to power rails of redundant cells are separated. This observation is further refined in this contribution. Accordingly, our proposed approach selectively inserts redundant MOSFET-stacks for the most vulnerable parts in a combinational design as a first step. Moreover, such additional stacks are switched off when the system operates appropriately and are switched on when they are required to improve performance or reliability.

A. Layout and Parameterization of the Redundant Stack

Due to the small area and power consumption of transmission gates, single MOSFETs have been chosen as the elements in our approach to dis- and reconnect the redundant stacks. By inserting pass transistors (subsequently called switch transistor) between the input net and the gate of a redundant transistor, it is possible to reduce the charge imposition on the redundant stack during the standby phase. A further transistor (called stack transistor) which will be also switched with the same signal as the switch transistors is positioned between the output net and the redundant stack to ensure a complete disconnection of the redundant parts and the original design during standby phase. Figure 1 exemplarily depicts an enhanced NAND2 cell with a redundant n-MOSFET stack. During standby phase, the switch and the stack transistors are turned off, which leads to an undefined but low voltage level at the gates of the redundant transistors. When the switch transistors are turned on again the associated voltage levels of the input nets are also applied to the gates of the redundant transistors. Hence, the output capacitance is discharged through the original as well as through the redundant transistors.

To select the best possible parameter set comprising transistor width and type (for both switch and stack transistors), a multitude of spice simulations were made. Thereby, the impact on timing, area and power had to be considered. For instance, inserting enhanced redundant transistor stacks increases the area and power consumption of the design. It also alters the timing in both phases due to a higher output capacitance of a driving cell that is connected to an enhanced cell. This results in rising transition and propagation delays for the driving cell. During standby phase the additional capacitance consists only of the drain capacitances of the switch transistors. When operating in the enhance phase though, the capacitance also comprises the input capacitance of the redundant transistors, which in turn adds to the output capacitance that the preceding cell has to drive. The difference is roughly a factor of two. On the other hand, the driving strength of the enhanced cell is doubled for the targeted transition direction, the one that is vulnerable to GOB.

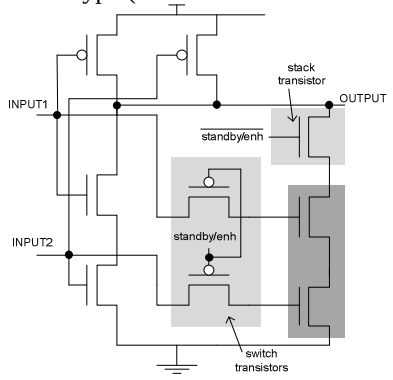

Figure 1 NAND2 cell with a redundant stack and enhanced elements (switch and stack transistors)

Furthermore, when a GOB occurs at a doubled transistor, the transition and propagation delay of the enhanced cell (when in enhance phase) is then much smaller as compared to the basic cell harmed by the same GOB. In this case, the overall transition time including both the driving cell and the enhanced cell is significantly smaller. Because of such major impact, all possible parameters were chosen with regard to the propagation delay from the input of a driving cell to the output of an enhanced cell where one transistor of a doubled stack is defective. Considering varying parameters like driving strength, additional input and output capacitances and the degree of the GOB, it was found that the redundant stack transistors should be of the same device type and width as the original transistors. A different case holds true for the switch transistors which need to be as small as possible. Such selections aim at best performance results of the design.

B. Insertion strategies

Another important aspect deals with the question which existing stack of the cells should be selected for the insertion of a redundant stack. First, defined limits for the probability P_{ON} have been set (0.5; 0.66). For clarification, one stack of every cell is doubled when $P_{ON} \geq 0.5$, thus, it only depends on whether the n-MOSFET or the p-MOSFET is turned on more often. By contrast, when $P_{ON} \geq 0.66$ only those cells are enhanced where at least one transistor is turned on for more than or equal to two-thirds of the whole operating life time. The second criterion is chosen as regards the overall propagation delay. With such a limit given, only cells are selected for the insertion that are part of critical paths. Here, critical means those paths where the propagation delay is equal to or slower than $p_D \cdot t_{P\_WC}$. t_{P_WC} is the overall worst-case propagation delay of the design and p_D is the chosen limit, called delay parameter, which can be 0, 0.75 or 0.9. For $p_D = 0$, all cells are selected and enhanced with redundant stacks. Contrary with $p_D = 0.9$ only cells are possibly modified that are found in paths where the propagation delay exceeds the value of $0.9 \cdot t_{P\_WC}$. The following nomenclature will be used for the different possible insertion scenarios:

TABLE I. PARAMETER SETTINGS FOR THE DIFFERENT SCENARIOS

Limits for	**Name**					
	Enh1	Enh2	Enh3	Enh4	Enh5	Enh6
P_{ON}	> 0.5	> 0.5	> 0.5	> 0.66	> 0.66	> 0.66
p_D	0	0.75	0.9	0	0.75	0.9

III. Experimental Results and Discussion

The proposed enhancements were used to modify the following reference designs: a 16 bit ripple carry adder (rca16), a 16 bit carry look-ahead adder (cla16) and two representative ISCAS designs (c432, c1908) whose original design parameters are displayed in the following:

TABLE II. Design parameters of the original designs

Design	Design Parameters		
	Overall transistor gate area [nm²]	Dynamic power consumption (10 Mhz) [nW]	Worst case delay [ns]
c432	25.8	5.4	0.616
rca16	43.5	11.6	0.445
cla16	55.9	12.3	0.306
c1908	60.1	14.1	0.917

The designs and the enhancements were implemented at transistor level with an industrial 65 nm ASIC library. To be able to simulate the mechanisms of GOB with Spice, we chose an equivalent circuit model from Renovell [9]. With this model it is possible to simulate a GOB on any horizontal and vertical location between gate and substrate, with an arbitrary resistance of the breakdown path.

Numerous simulations were made to investigate the behavior of the designs with and without the enhancements. In order to simulate wearout, the insertion of failures is based on the Weibull cumulative distribution function [10]:

$$F(t) = 1 - e^{-\left(\frac{t}{ttf}\right)^{\beta}} \quad (3)$$

with β as the Weibull parameter, t is the time and *ttf* (time-to-failure) as the characteristical operating lifetime with a failure probability of 63 %. The *ttf* was varied for every transistor according to equations (2). Finally, a uniformly distributed random number between 0 and 1 was assigned to each *F(t)* of a transistor. Hence, by using equation (3), a dedicated time $t = t_{BD}$ could be calculated for every transistor, which is the point in time where GOB occurs. Then, the breakdown current I_{BD} through the defective gate obey an exponential increase over time [11]. Based on the described failure insertion, it is possible to assign a realistic and individual defect development for every transistor. Note that the time steps that are depicted in the following are relative and dimensionless and a GOB can occur for every time $t > 0$. Concerning the simulation termination and reliability analysis, an entire design was considered erroneous when it produced wrong logical output signals.

A. Design Overhead

As redundancy is inserted into the original designs, area and power consumption are increasing. Compared to the according reference designs, it is not surprising that the designs for Enh1 (where the redundant stacks are inserted independent of P_{ON} and the overall propagation delay) exhibit the highest area overhead with an increase of more than 60 % and less than 75 % due to the switch and stack transistors. Furthermore, the area of the designs for Enh4 is roughly at 140 % due to the delay consideration. By contrast, for the other enhanced types (Enh2, Enh3, Enh5, Enh6), the area overhead is mostly below 120 % compared to the original designs due to the enhancement restrictions explained in section II B.

The relative overhead of dynamic power consumption is at worst (Enh1 of cla16) 12 % higher than the original one during the standby phase. However, during the enhance phase the power consumption can be more than three times higher compared to the original one. However, first of all the standby phase is by far the common mode of operation. Second, the original design does not properly operate at all anymore when the enhanced design changes into enhance phase. Thus, it is recommended to switch to the enhance phase only when defects have already occurred, because the overhead of power consumption by than is small in contrast to other design constraints. Furthermore, with the passing of operation time and occurrences of gate oxide defects the leakage current through the defective transistor gates becomes so dominant that the power consumption overhead due to the enhanced modules is irrelevant. Further on, the impact of the higher capacitive loads in the design (section II A) is summed up in the following table where the relative overhead of the propagation delay for the standby phase is depicted. It needs to be noted that although the initial timing is slower, the timing of the enhanced designs remains nearly constant during wearout. Hence, after a certain lifetime, the timing is actually better than in the reference design and timing requirements can longer be met (see also section III D).

TABLE III. Relative Worst-Case Delay Overhead

Design	Insertion scenario					
	Enh1	Enh2	Enh3	Enh4	Enh5	Enh6
c432	1.137	1.0693	1.0693	1.1251	1.0686	1.0686
rca16	1.1322	1.0828	1.0831	1.0943	1.0831	1.0828
cla16	1.1228	1.1041	1.1032	1.1107	1.0950	1.0950
c1908	1.1369	1.0244	1.0186	1.0955	1.0243	1.0198

B. Reliability Improvements

One major objective of our approach is the improvement of reliability during wearout. The reliability *R(t)* of a component itself is the probability of the device to perform as intended until time t [10]. The performed simulations have shown increased reliability for every insertion scenario and design due to the redundant transistor stacks, as these maintain the voltage levels on the affected nets when the stacks are switched on. A plot for the design rca16 is exemplarily depicted for all designs in Figure 2. Here, the reliability of the reference design is displayed as a dotted line. Both insertion scenarios Enh1 and Enh4 (which have most redundant stacks) reach the best result. Translated into time, some of the designs were able to correctly operate two times longer than the original design. For example, at time $t = 4.5$ (where no original design works properly anymore), one third of the Enh1 designs

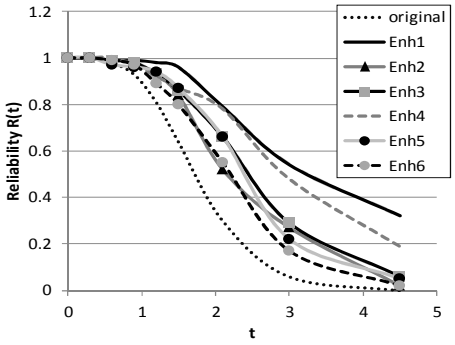

Figure 2 Reliability of the original and the insertion scenarios of rca16

still provide correct outputs, whereas also the design scenarios Enh4 exhibit noticeable robustness against GOB. The other insertion scenarios are less reliable but still significantly better than the reference designs. To compare the improvements in a quantitative manner, the Mean Time To Failure (*MTTF*) has been calculated for every design which is depicted in table IV. Here, the *MTTF* is the average time a design operates until it fails, which is equal to the expected lifetime (as the design cannot be repaired) [10].

TABLE IV. RELATIVE MTTF IMPROVEMENTS

Design	Insertion scenario					
	Enh1	Enh2	Enh3	Enh4	Enh5	Enh6
c432	1.34	1.06	1.13	1.21	1.12	1.09
rca16	1.76	1.31	1.41	1.63	1.35	1.24
cla16	1.72	1.10	1.10	1.48	1.14	1.10
c1908	1.63	1.03	1.03	1.65	1.12	1.18

Due to the most redundant stacks, the Enh1 designs perform best on average, followed by the Enh4 designs. *MTTF* improvements of up to 76 % can be achieved when all cells are at least partially enhanced with redundant stacks (Enh1). Moreover, when trying to keep the area overhead small (Enh2, Enh3, Enh5, Enh6), the *MTTF* still improves between 5 % and 20 %. In between the designs for Enh4 are found, which is a compromise of the other two cases.

C. Improvements of the Propagation Delay

The relatively small improvements for the c432 designs, as regards *MTTF*, correspond to the fact that the GOBs rather impair the propagation delay than they produce logic failures. To further detail this effect, the average delay for numerous simulation runs is depicted in Figure 3 for the insertion scenarios and the original designs (dotted line). Thus, at $t = 1.5$ the delay of the original designs is almost twice as high. This in turn means, that if an arbitrary constraint would be exist for the clock signal at $t_{CLK} = 1$ ns, more than 50 % of the original designs would already fail at time $t = 1$ due to timing issues. By contrast, all enhanced insertion scenarios provide better propagation delays than the original designs over time (except Enh2). Referring to the previous example with the clocking constraint, the designs for Enh1, Enh3, Enh4 would still satisfy the timing until $t = 1.2$ (the designs for Enh4 actually until $t = 1.5$). Hence, this analysis clearly states that additional *MTTF* improvements exist when timing requirements are considered as well. Similar conclusions can also be drawn for the designs that have more balanced paths, whereas not all designs achieve such a significant delay advantage over time. Figure 4 depicts the development of the propagation delay for rca16 (Enh6) in detail. Here, the delay for the Enh6 insertion scenario is related to the propagation delay of the original design. The collected values are shown as a boxplot diagram. That means, the boxplot divides the results into four blocks, whereas every block represents one fourth of all values (white box = lower quartile; grey box = upper quartile) and the median at 50 % is between the grey and white boxes. The ends of the markers represent the minimums and maximums of the results. Even though, the shown design rca16 does not perform best (compared to the other designs), it can still clearly slow down the delay degradation compared to the original design when the proposed enhancement are implemented and switched on. The median decreases with every further time step. From time $t = 2.1$ on, almost 75 % of the working designs are faster than their original counterparts, and at $t = 3$ almost all designs are faster.

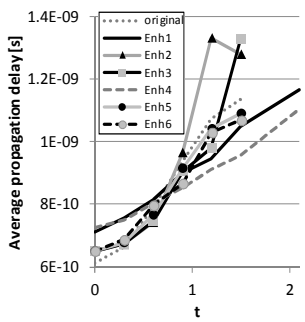

Figure 3 Average propagation delay of every insertion scenario (c432)

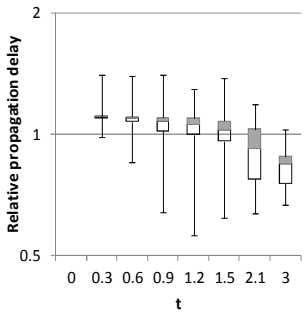

Figure 4 Relative worst case propagation delay of Enh6 (rca16)

IV. CONCLUSION

This contribution focuses on improvements of lifetime reliability. Therefore, four different combinational designs were enhanced with redundant transistor stacks, which double the stacks of a given, industrial standard cell library. The stacks were implemented, such that the redundant stack is electrically separated from the original stack. Moreover, pass transistors were used to dis-/reconnect the added redundancy in dependence of application needs. The implemented insertion scenarios clearly outperform the reference designs in terms of GOB reliability, whereas the *MTTF* improvements range from 5 % up to 75 %. The design that applied the most redundant transistor stacks (overhead: +60 % area, +12 % power) also achieved the best improvements. A cost-effective alternative is the insertion scenario Enh4, which adds an area overhead of 25 % up to 45 % while providing reliability improvements of up to 48 %. Moreover, when additional timing constraints of the clock are considered, the insertion scenario Enh4 provides the best results as regards the timing degradation which translates into additional improvements of lifetime reliability. The other insertion strategies appear as an alternative only if area requirements are very restrictive. Lastly, the use of redundant stacks applied to standard cells provides a possibility for design improvements based on standard ASIC cell libraries. However, the additional redundancy also impairs area, power consumption and propagation delay during common operation mode, so that a tradeoff has to reflect the application needs.

REFERENCES

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

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

[3] Renovell, M., Gallière, J., Azaïs, F., Bertrand, Y.: "Delay Testing of MOS Transistor with Gate Oxide Short", In Proc. of ATS, 2003.

[4] SIA, "International Technology Roadmap for Semiconductors", 2007.

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

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

[7] Saemrow, H.etal.: "Twin Logic Gates - Improved Logic Reliability by Redundancy concerning GOB", In Proc.of SBCCI, 2009.

[8] Xiaojun Li: "Deep Submicron CMOS VLSI Circuit Reliability modeling, simulation and design", Dissertation, 2005.

[9] Renovell, et al.: "Modeling the Random Parameters Effects in a Non-Split Model of GOS", In Journal Electronic Testing, 2003.

[10] Koren, I..: "Fault-Tolerant Systems," Morgan-Kaufmann, 2007.

[11] Lindner, B. et al.: "Growth and scaling of Oxide Conduction after Breakdown", In Reliability Physics Symposium Proc., 2003.