On Reliable Circuits and Systems: How Reliability Considerations Are Reshaping Oxide Scaling, Device Geometry, and VLSI Algorithm

M. Alam, H. Kulluoglu, B. Paul, K. Kang, and K. Roy
School of Electrical and Computer Engineering
Purdue University, West Lafayette, IN 47906
TEL: +1-765-494-5988 email: alam@purdue.edu

1. Introduction

The reliability issues can broadly be divided into two groups: initial, process-related reliability that defines manufacturing yield of an IC (e.g., Na\textsuperscript+ related bias-temperature instability in 1960s, CMP induced shorts between interconnect metal lines, junction short due to metal spiking, etc.\cite{1}) and the long-term, intrinsic device-operation related reliability (e.g., oxide breakdown, hot carrier degradation, radiation-induced damage, etc.\cite{2}) that defines IC operation during product lifetime. From the early days of semiconductor industry, the initial time-zero reliability has determined process options, and the long-time reliability has defined device design parameters. Indeed, IC designers have routinely accounted for manufacturing variation (parameterized in $C_\text{q}$ numbers) and device degradation-induced parameter shifts by requiring that IC performance is achieved even with a certain amount of parameter variation (typically ±15% of the nominal values). However, for sub-100 nm CMOS technology, reduced long-term reliability budget and increased power-dissipation have dramatically restricted device design options (Sec. 2 and 3) and therefore, the aforementioned worst-case, time-independent, guard-band limited design may often be too conservative for practical design and must be supplemented by a reliability-aware VLSI design algorithm (Sec. 4). In this paper, we briefly describe the evolving role of reliability considerations in process-device-circuit-architecture hierarchy of IC design.

2. Gate Dielectric Reliability: Beginning of Circuit-Specific Reliable Design

Like Negative Bias Temperature Instability (NBTI) problem in 1970s\cite{3,4}, the Hot Carrier Degradation (HCD) problem in 1980s\cite{5,6}, the defining reliability problem of 1990s was gate oxide reliability (also called Time Dependent Dielectric Breakdown or TDBB)\cite{7-12}. Unlike the NBTI problem, which disappeared after widespread adoption of NMOS and buried channel PMOS devices, unlike the HCI problem which was managed by a combination of LDD and redistribution in supply voltage, TDBB turned out to be a more difficult reliability issue because reduction in TDBB lifetime is directly related to statistical failure distribution of ultra-thin oxides as oxide thickness (1-5 nm) becomes comparable to the defect dimension (~1 nm). Simply put, the time to oxide breakdown is $T_{\text{BD}} = N_{\text{D}} R(V_{\text{BD}})$, where $N_{\text{D}}$ is the number of defects required for breakdown and $R(V_{\text{BD}})$ is the supply voltage dependent defect generation rate. In order to increase performance, gate oxides had to be scaled faster than supply voltage (see Fig. 1,\cite{13}). Thinner oxides require fewer defects to complete the defect chain to short-circuit the channel to the gate (thereby causing transistor failure). Although probability of stacking of defects on top of each other in any given transistor is extremely unlikely -- after all defect generation in oxides is statistically independent and there are many other sites for the defects to occupy...
in the bulk of the oxide — however, given that an IC has millions of transistors, the probability of formation of such path in at least one of the transistors is a statistical certainty. With geometry and statistics against gate oxide scaling, device designers must reduce operating voltage. \( V_{DD} \), so that trap generation rate, \( R(V_{DD}) \), is reduced and \( T_{DD} \) is prolonged. Despite some lingering debates, the broad consensus is that required reduction in supply voltage for NMOS transistor is small enough so that the performance metrics (as defined by SIA-ITRS roadmap, see Fig. 1) can still be met down to 1.0 nm at 1.0-1.3 volts [11,14,17]. However, the reduction of supply voltage required to ensure traditional PMOS reliability would result in unacceptable circuit performance [14,15].

This is precisely where the device-circuit co-design came to the rescue of IC development. In a remarkably insightful paper, B. Weir from Bell Labs [16] discussed the possibility that, with decreasing supply voltage, the power dissipation through the first percolation path may be too small to cause transistor failure. Her systematic study of transistor parameters before and after “breakdown” (as a function of oxide thickness, supply voltage, etc.) supported this hypothesis [16-19]. Since then formal theories [20, 21] have quantified the improvement of reliability associated with such “soft breakdown” and have demonstrated geometrical improvement in IC lifetime if a transistor can “survive” the first soft breakdown. This enhanced lifetime can then be traded for higher operating voltage required to achieve IC performance metrics (arrows in Fig. 1) [18,20,21]. However, the requirement that soft-breakdown does not destroy transistor operation must be supported by a case-by-case study, because requirements of transistors performance vary with circuits. Various groups have investigated the effect of first breakdown on digital logic [22], memory elements [23], and analog blocks [24] with results that have generally supported Weir’s original hypothesis that soft-breakdown does not cause transistor failure. Although it is impossible to test all circuits exhaustively for tolerance to soft-breakdown, simple arguments based on Bayesian statistics can ensure that the overall IC reliability margin is maintained. For example, if major blocks in a microprocessor (e.g., memory, logic) can be made insensitive to soft-breakdown, then the probability that the remaining blocks (e.g., analog block) experiencing soft-breakdown can be made exponentially small by restricting is footprint to specific values (see Fig. 2, [25]). Moreover, design techniques like inclusion of parity-bits, already implemented to protect against radiation-induced damage [47,48] can also act as a redundancy bit for TDDB breakdown. Such reliability-aware co-circuit design will be norm in the future.

Although Fig. 1 suggests that oxide could be scaled to 1.0 nm and still meet performance targets, gate leakage increases exponentially with reduction in oxide thickness, therefore in practice gate oxide scaling must necessarily slow. Moreover, many people believe that performance goals and dopant fluctuation considerations dictate that the supply voltage remain close to 1.0 volts (related to Silicon bandgap). With gate oxide thickness and supply voltage approximately constant, the TDDB issues becomes a purely circuit-design consideration, as discussed above.

3. Negative Bias Temperature Instability and Geometry of Degradation

With gate oxide thickness and voltages essentially fixed, one might wonder if all reliability concerns are simultaneously resolved. Unfortunately, this is not the case! In order to increase performance without correspondingly scaling of oxide thickness or supply voltage, one must either reduce parasitic capacitance or increase channel mobility. Similarly, to increase the On/Off current ratio for better power-management, one needs better gate electrostatic control. There are several choices: replace \( \text{SiO}_2 \) with high-k gate dielectric, replace channel Si with GeSi or strained Si, use multiple gate devices with back-gate biasing options or use single surround gate devices [26-31]. The reliability issues for high-k gate dielectrics and
strained channel material are still evolving, so it is difficult to analyze their circuit implications. However, the remaining two options are better understood and are discussed below.

Among the surround gate devices, partial or fully depleted devices like VRG [28] or FINFETs (also known as X-FETs or Omega-FETs) [29-31] are promising options. Unfortunately, the change in channel geometry enhances NBTI. NBTI was a serious reliability problem during the early days of semiconductor industry while PMOS-only circuits dominated the technology landscape. It reappeared in 1990s when surface channel PMOS (in favor of buried channel PMOS) were reintroduced for high-speed I/O circuits, without corresponding reduction in supply voltage [32]. As electric field continued to increase with disproportionate oxide scaling compared to supply voltage (oxide field has increased almost ten-times over last ten years), NBTI has become an important reliability issue for logic CMOS transistors as well [33-36]. And more recently, the use of Nitrogen in oxides (to reduce gate leakage) has further enhanced NBTI degradation.

To appreciate this geometry dependence, consider the mechanics of NBTI degradation: inversion layer holes in PMOS are captured by Si-I-I hands at the Si/SiO$_2$ interface through field assisted tunneling. This weakens the Si-I-I bonds, which are then more easily broken at high temperature. The H diffuses away leaving behind an unpassivated Si bond which acts as trapping center ($N_T$). These interface traps ($N_T$) increase threshold voltage and decrease channel mobility, with overall degradation in transistor performance. Interestingly, detailed calculation show that a short time after the initiation of the NBTI degradation, further increase in $N_T$ is dictated not by the rate of bond-dissociation (because Si traps can easily be annealed by nearby free H), but by the rate of H diffusion away from the interface and it given by $N_T$ rate, where A is constant. i is time, and $n=1/4$.

The diffusion of H is given by $D_H \nabla^2 N_T = 0 (DH$ is the diffusion coefficient, and $N_T$ is the hydrogen concentration) which is similar to Poisson equation that controls the electrostatic potential $\phi$ in the channel, i.e. $K_e \nabla^2 \phi = 0 (K_e$ is the dielectric constant). Therefore, the same surround gate geometry that provides better electrostatic control also allows faster out-diffusion of hydrogen (or increased n, see Fig. 3) and hence increase in NBTI degradation compared to planar devices [39-42].

Once again we have a circuit/device co-design issue: from the circuit perspective, if the oxides and supply voltages can not be scaled, a device designer must scale the body through surround gate approaches [26-31]. But this, as we now realize, must come at a price of increased NBTI degradation. Therefore, a co-optimization is the only practical approach available for design of these circuits.

The other option frequently mentioned for better electrostatic control, especially for low power designs, involves circuits with backgate bias. In these devices, the backgate is used to statically modify the global threshold to increase in circuit speed or dynamically modify local threshold to reduce power-dissipation. One might anticipate that such modification at substrate bias would not change oxide reliability, but this assumption is incorrect. For PMOS transistors with gate injection, even a moderate body bias can substantially change...
the ratio of soft to hard breakdown [43]. As such, the guidance of device design in Fig. 1 and discussed in Sec. 2, based on the theory of soft-breakdown, may no longer be appropriate. Therefore, it is important that every new circuit design proposal is cross-checked for reliability implications [43] to avoid unanticipated problems.

4. Reliability-Aware Statistical VLSI Design

Given the device and circuit-specific reliability issues discussed above, it is obvious that standard VLSI design must be modified to account for reliability-aware design. Research in the area of VLSI circuit design has traditionally focused on the trade-off among area, performance, testability, and power dissipation. The key feature of such deterministic optimization is that it involves analysis of a static and uniform network of transistor and interconnect nodes in response to a set of time-dependent test-patterns. Such analysis ignores the fact that the transistors parameters are not time-independent, but rather they change with node activity, input patterns, and biasing configuration, as discussed in Sec. 2 and 3. Moreover, transistor parameters are not spatially uniform, but vary significantly from die to die and from wafer to wafer due to process-variability. As we mentioned earlier, traditional VLSI design bypasses the analysis and optimization of such dynamic, nonuniform network by approximating this problem into optimization of a uniform static network with certain guard band. This allows the optimized static circuit to continue functioning even if the devices have randomly distributed parameters which can become faster or slower over a period of time. Such worst-case guard-band limited VLSI design, while routine and practical, is inefficient and involves considerable penalty in the area-performance-power budget. As technology scales to sub-100 nm regime, such conservative design approaches can not be sustained because the area-performance-power budget is shrinking fast, and the process-induced variation and time-dependent shifts in transistor parameters are increasing rapidly. Moreover, as discussed in Sec. 3, these parameter variations are rapidly approaching intrinsic limits which can not be reduced by process-modifications without sacrificing other performance metrics.

Therefore, the question is: “How would one optimize a VLSI circuit whose underlying network contains transistors with statistically distributed parameters that themselves evolve in time and retain the cumulative memory of its past states?” One obvious choice in dealing with problem is based on statistical design methodology [44] based transistor sizing and activity re-balancing to address these design constraints and to quantify their performance implications compared to traditional guard-band limited design approaches. In short, the statistical methodology would change the design paradigm of VLSI circuits by adding a fifth variable – that of statistically distributed device parameters and time-evolution of dynamic network – to the traditional optimization of performance, area, testability, and power dissipation.

As an example of this design process, consider activity dependent optimization of ISCAS benchmark circuit C432 that contains several hundred transistors (taken from [44]). Assuming the circuit delay and area being two global variables the circuit must be optimized for, one can perform the transistor sizing optimization to reduce NBTI induced degradation. Simply put, if NBTI degradation is 15% of Vth, one could increase the width of all the transistor by 15%, thereby pre-biasing the drive current with 15% overhead so that even after NBTI degradation, the circuit will remain fully functional. On the other extreme, we could stay with current area, but then accept 15% increase in delay. However, the optimized solution will involve increasing the width of transistors that are most NBTI-active and reduce the area of the transistor which are not in NBTI-critical path such that the area-delay product is minimized (see Fig. 4). The Lagrange multiplier based algorithm [45] is general enough for routine use in more complex circuits. Details of the algorithm and its application...
for wide class of circuits will be published elsewhere [46].

5. Conclusions:

In this paper, we have outlined the growing importance of reliability-aware circuit design. Such reliability-aware device design has long been used for radiation-hard applications [47,48]. Soon, however, such design techniques may become essential for any IC design based on newer transistor geometries of sub-100 nm CMOS technology. In addition to establishing the importance of device-circuit co-design, we also have proposed novel use of statistical design techniques based on transistor sizing and activity rebalancing for design with transistors with significant time-dependent parameter variation.

Acknowledgement:

The authors acknowledge financial support from NSF/SRC and from Birk Nano-technology Center at Purdue University, West Lafayette, IN.

References


