ReverseAge: An Online NBTI Combating Technique using Time Borrowing

Seyab Khan          Said Hamdioui  
Computer Engineering Laboratory  
Delft University of Technology  
Mekelweg 4, 2628 CD Delft, The Netherlands  
E-mail: {M.S.K.Seyab, S.Hamdioui}@tudelft.nl

Abstract—As semiconductor manufacturing has entered into the nanoscale era, Negative Bias Temperature Instability (NBTI) has become one of the most significant aging mechanisms leading to reliability issues. This paper presents ReverseAge, a technique that detects delay due to NBTI and utilizes design timing margins to ensure reliable circuit operation. First, it presents a scheme to detect the NBTI induced delay. Second, it presents a technique to tolerate the errors; the technique exploits the available design timing margins to compensate for the NBTI induced delay. The evaluation of ReverseAge has been performed by integrating it in an ISCAS-89 benchmark circuit. The simulation results show 3.77% improvement comes at the cost of 3.77% area and 1.4% power overheads.

I. INTRODUCTION

With the relentless pursuit for smaller CMOS process technologies, timing errors due to aging mechanisms have become a limiting factor for transistor/circuit reliability [1]. Industrial data reveal that as oxide thickness reached 4nm, Negative Bias Temperature Instability (NBTI) has become one of the most significant aging mechanisms [2,3]. NBTI degrades the performance of a PMOS transistor under a negative gate stress at elevated temperature [4,5]. The effects of NBTI on a PMOS transistor include: (a) threshold voltage increment, (b) drain current degradation, and (c) delay increment [4,6]. These effects show up themselves at the circuit level by increasing circuit delays and in turn circuit timing errors.

In recent years, there is an escalation of interest in two aspects of NBTI: (a) modeling and monitoring, and (b) mitigation and design for reliability. NBTI modeling has been done both at the transistor [4,5,7,8] as well as at the circuit [6,9,10] levels. The NBTI monitoring schemes proposed so far can be classified as follows.

• Transistor parameter monitoring: These schemes are based on monitoring the parameters of some selected transistors to measure the NBTI impact on the circuit. For instance, Denias et al. in [11] monitored the transistor drain current, threshold voltage, and transconductance to quantify the impact. Kang et al. in [12] monitored the leakage current to characterize NBTI.

• Circuit parameter monitoring: These schemes are based on monitoring the circuit parameters to measure the NBTI impact. For instance, Kim et al. in [13] monitored the beat frequency of two ring oscillators (which is a special structure used on the chip), one stressed and the other unstressed, to characterize the impact. Agarwal et al. in [14] monitored NBTI using a stability checker circuit. Keane et al. in [15] characterized NBTI in terms of the control voltage of the two delay locked loops.

Although all the above proposed schemes measure NBTI impact, they require significant area overhead and/or do not give realistic measure of NBTI impact. For example, transistor parameters monitoring schemes [11,12] require enormous area overhead when it is applied to a larger group of PMOS transistors that degrade at different rates. Similarly, the circuit parameter monitoring schemes [13,15–17] are not realistic because the monitoring circuits are activated for a selected period of time. Therefore, a scheme that incurs lower area overhead and realistically measure NBTI impact on the circuit is required; this is one of the topics addressed in this paper.

Apart from NBTI monitoring, researchers have focused on NBTI mitigation and design for reliability techniques. Research has been conducted to tolerate NBTI impacts at transistor/circuit levels [4,9,18]. The techniques proposed so far can be classified as:

• Pre-manufacturing techniques: These techniques are based on the process optimization or design optimization. Scarpa et al. in [19] suggested several process optimization techniques (e.g., oxide nitridation, adding fluorine to gate oxide etc.) and Islam et al. in [18] proposed a uniaxial strained silicon transistor (with SiGe source/drain and contact etch stop layer) to mitigate NBTI degradation. Similarly, Paul et al. in [9] suggested 8.7% gate oversizing at the design stage to mitigate NBTI.

• Post-manufacturing techniques: These techniques are based on the contents flipping, input vectors control and parameters tuning. Kumar et al. in [20] proposed a technique to flip the contents of SRAM cells to mitigate the NBTI impact. Wang et al. in [21] proposed a technique to control the input vectors of a node during its standby mode to reduce the probability of stress and hence NBTI. Wang et al. in [10] suggested a set of techniques including: $V_{th}$ tuning, $V_{dd}$ tuning and temperature reduction to compensate for the NBTI impacts.
have not extended their analysis to deal with NBTI at a higher level. Alam et al. in [4] have modeled NBTI at the device level and presented the overall dynamics of NBTI as a reaction diffusion process. The model is usable at a higher level such as circuit level. Since in this work, NBTI analysis is done at the circuit level, model of [4] will be used that relates $N_{IT}$ with time ($t$) as follows:

$$ \Delta V_{th}(t) = \left( \frac{k_f N_o}{k_f} \right)^{2/3} \left( \frac{k_h}{k_{H_2}} \right)^{1/3} (6D_{H_2}t)^{1/6}, $$

where $N_o$, $k_f$, $k_h$, and $D_{H_2}$ represent initial bond density, H to H$_2$ conversion rate, H$_2$ to H conversion rate, and H$_2$ diffusion rate inside SiO$_2$ layer, respectively.

Interface traps are assumed to be positive charges remaining at the Si-SiO$_2$ interface that oppose the applied gate stress resulting in the threshold voltage increment ($\Delta V_{th}$). The relation between $N_{IT}$ and $\Delta V_{th}$ is given by [9]:

$$ \Delta V_{th} = (1 + m)qN_{IT}/C_{ox}, $$

where $m$, $q$, and $C_{ox}$ are the holes mobility degradation that contribute to the $V_{th}$ increment [6], electron charge, and oxide capacitance, respectively. On the other hand, NBTI annealing takes place under the positive gate stress; in this case, the H atoms anneal back towards the Si-SiO$_2$ interface as shown in Fig. 1(b). The H atoms/molecules anneal the $\equiv$Si- bonds at Si-SiO$_2$ interface resulting in lower $N_{IT}$ and hence lower $\Delta V_{th}$.

The NBTI induced $\Delta V_{th}$ of PMOS transistors causes additional gate delay which is given by [9]:

$$ \Delta D = \gamma \frac{n \Delta V_{th}}{(V_{gs}-V_{th})} $$

where $n$ is the velocity saturation index and $\gamma$ represents the distribution of stress and recovery -positive and negative gate stresses- durations. The $\gamma$ dependence of the additional delay reveals that transistors in circuit having different stress/recovery duration will suffer from different delays, resulting in the circuit delay variation.

A simple CMOS inverter was synthesized using 45nm Predictive Technology Model (PTM) transistor models [30] and simulated using HSPICE for 10 years operation. To focus on NBTI, it is assumed that process variations and the other failure mechanisms, e.g. Hot Carrier Degradation, Electro-migrations and Time Dependent Dielectric Breakdown are not affecting the inverter. Throughout the simulation, $\gamma=50\%$ and NBTI parameters used to get $\Delta V_{th}$ are, $k_f=8 \times 10^{-4}$s$^{-1}$, $k_r=3 \times 10^{-18}$cm$^3$s$^{-1}$, $N_o=5 \times 10^{16}$cm$^2$, $D_{H_2}=4 \times 10^{-21}$cm$^2$s$^{-1}$ and $T=125^\circ$C [8]. Fig. 2(a) gives the $\Delta V_{th}$ increment of PMOS transistor due to NBTI; it shows that $\Delta V_{th}$ approaches 47.81mV after 10 years of operation. The curve follows 1/6 trend and have a good match with the experimental results presented in [3]. Fig. 2(b) shows $\Delta D$ i.e. the percentage increment of the inverter delay that approaches 12.37%. Paul et al. in [9] has suggested the same $\Delta D$ trend due to NBTI for the other gates of a circuit.
III. REVERSEAGE MOTIVATION

This section analyzes the distribution of NBTI impact in a circuit and explores the available timing margins of the circuits for tolerating NBTI. Usually in circuits there are a number of combinational logic clusters (CLCs) along with the sequential elements (e.g. FF’s). Fig. 3 presents a conceptual representation of a synthetic circuit with four stages (i.e. four CLC’s and five FF’s); the circuit is sufficient enough to illustrate the distribution of NBTI and the available timing margins. All CLC’s are customized such that they have nearly the same critical path delay just like an industrial design that balance all the stages as close as possible to reduce power and area of the design.

First, simulation has been done to explore NBTI induced delay distribution. Fig. 4 summarizes the delay distribution of the four stages. Three input sets -set1, set2, and set3- are considered. Each set of the four bars for each input set represents NBTI induced delay in the four stages. The figure shows that the input signals which causes enough delay in one stage of the circuit produce far less delays in the neighboring stages. For example, the input -set3- causes 19.09% additional delay in stage C1355 while it produce only 13.88% delay in stage C1908.

Second, the maximum allowable delay of a stage is defined for the expected worst operating conditions (e.g. high temperature and voltage droop) [25]. However, under normal operating conditions the parameter is unnecessarily strict and the setup margin (discussed in the next section) available will never be used [27]. The margin can be used to tolerate delays due to aging or process variations. The black portion of each bar in Fig. 4 represents the available margin in each stage under the given input set. For example, stage C1908 suffers from only 19.74% delay under input -set2- and have about 1.35% setup margin available -with respect to the worst effected stage C880- for the preceding stages. The setup margin can be used by the preceeding stage without causing any error in the circuit.

To summarize the distribution of NBTI and the available timing margin, it is argued that NBTI in a stage can be tolerated by borrowing time from the succeeding stage. The time borrowing will not affect timings of the next stage (details will be discussed in the next section). However, if NBTI causes timing errors in multiple adjacent stages then the time borrowing from the next stage will not be effective and may cause timing error. Under such conditions the timing error is tolerated by dynamically adjusting the clock frequency at system level [25].

IV. REVERSEAGE: ARCHITECTURE

This section describes an NBTI monitoring scheme along with a tolerating technique. The monitoring scheme detects the NBTI induced delay and produces an error signal. Circuitry of the error tolerance responds to the error signal and borrows time from the succeeding stage. After lending time to preceding stage, the succeeding stage operates in its own timing limits.

A. NBTI Monitoring

Fig. 5 shows an example circuit to illustrate the NBTI monitoring concept. The monitor is augmented to the flip-flop (FF) that is connected to the combinational logic cluster (CLC) output (Cout) as shown in Fig. 5(a). Without considering NBTI, Cout approaches FF2 well before its setup margin that results in sensor output is low as shown in Fig. 5(b). However, with NBTI, delay of the CLC increases significantly and Cout creep into the setup margin. It is important to not that the FF2 will still capture correct value at rising edge of the clock (due to the available setup needed). However, the monitor predicts the setup time violation and produce an error signal as shown in Fig. 5(c).
in Fig. 5(c). The scheme predicts- NBTI in the circuit and has lower overhead as compared to Razor [28] and Canary logic [16] that has shadow FF and canary FF augmented to the normal FF, respectively.

Fig. 6 shows the implementation and Fig. 7 shows signal waveforms at the internal nodes of the monitor. The monitor consists of an NBTI Delay Detector (NDD), a C-element, and a keeper circuit. NDD detects the NBTI in the CLC in a transient form, C-element sensitize the transient and then the keeper circuit converts it to a stable signal. The NDD design is simple that consists of two delay elements (A and B) and a three inputs AND gate (C). The CLC output (C_{out}) and Clk are inputs to the delay elements A and B, respectively. Without considering NBTI in the CLC, C_{out} arrives earlier than the clock (Clk) edge and FF output (Q). Element A delays and inverts C_{out} (C_{outA}) while B delays the Clk (ClkB) such that ClkB goes low before C_{outA} and Q goes high, as a result output of the gate C remains low as shown in Fig. 7(a). However, by considering NBTI in the CLC, the output C_{out} will be delayed and there will be an interval when all the three inputs (C_{outA}, ClkB nad Q) of the AND gate are high that results in high output N_{out} as shown in Fig. 7(b).

The NDD output signal N_{out} is high for a small duration, a transient with a width equivalent to the NBTI induced delay in the CLC. The output of NDD may or may not activate the NBTI tolerating circuitry (discussed in the next subsection). The C-element and the keeper circuit shown in Fig. 6 are similar to the one presented [27]. The C-element senses transient output of the sensor during low as well as high clock cycles. Sensitivity of the C-element is adjusted so that it is turned on by NBTI induced delays and have a very small response time. Fig. 7(c) shows that N_{outA} arrives the C-element at the beginning of the clock high. Output of the C-element is input to the keeper circuit of Fig. 6. The keeper circuit comprises cross-coupled inverters that are turned on by the momentary on state of the C-element and maintains a stable output (K_{out}) signal.

The monitor design is verified using HSPICE simulation. An inverter chain that act as CLC, FF’s and the monitor circuit are synthesized using 45nm PTM transistor models [30]. Transistors in the monitor are optimally sized to have smaller response time (40ps). Fig. 8 shows the CLC output (C_{out}), Clk and the monitor output (K_{out}) signals. The figure shows that without considering NBTI in the CLC, the C_{out} approaches well before the Clk (about 300psec). Under this condition C_{out} is beyond the FF setup margin as a result the monitor output remains low. However, by considering NBTI and after 10 years of operation, C_{out} and Clk approaches the monitor/FF nearly at the same time (with only 29psec time difference). The monitor sensitize the setup time violation of the FF and produce a high K_{out} signal.

B. NBTI Tolerance

Once the NBTI induced delay is detected by the monitor circuit, the next step is to tolerate it by utilizing design timing margin of the circuit. In ReverseAge, first the possibility of utilizing time margin of flip flops in succeeding or preceding stages of the circuit is explored and then a technique to tolerate NBTI is proposed.

Delay of a stage is determined by delays of the CLC and adjacent FF’s. To be in timing limits delays of the CLC and FF should be less than clock period and the input signal must pass through the stage in one cycle. Fig. 9(a) shows a stage delay with the maximum allowable delays in the CLC and FF of the stage. The figure shows that delay in the CLC is limited by hold and setup times constraints of the the adjacent FF’s. Relaxing either the hold time of the preceding FF or setup time of the succeeding FF increases the maximum allowed delay to the CLC. However, relaxing the hold time of the preceding FF requires complex clocking schemes and implies significant performance loss [27]. Therefore, the delay portion of the CLC is increased by relaxing the setup time of the succeeding FF i.e. shifting the rising edge later as shown in
Fig. 9. A stage delay with (a) maximum allowable delay of the CLC and FF, (b) Time stealing by a CLC using setup margin of FF in the next stage

Fig. 10. Schematic of Time stealing technique: (a) Master slave flip flop with a Specialized Clocking Scheme, (b) Specialized Clocking Scheme for the master latch

V. CASE STUDY AND ANALYSIS

A five stage 4×4 add shift multiplier circuit (S349 in ISCAS-89 benchmarks) is taken as a case study to illustrate effectiveness of ReverseAge technique. Predictive Technology Model (PTM) cards of 45nm technology nodes are used to synthesize the circuit. The clock (Clk) used is 10ns (Clk=0.1GHz) with 50% duty cycle. NBTI induced delay degradation is injected into the circuit by using a Verilog-A module under the conditions (V_{DD}=1.0V and T=125°C). First, the circuit is simulated for 10 years operation to get temporal degradation due to NBTI degradation. The simulation identified five critical flip-flops among the total 15 FF’s of the circuit. Thereafter, proposed elements-monitor and toleranting circuitry- of the ReverseAge are integrated into the benchmark.

Effectiveness of the ReverseAge along with the area and power overheads are discussed below.

ReverseAge Effectiveness

The effectiveness of ReverseAge is illustrated by varying the NBTI induced delay in the CLC and observing the error signal in the monitors attached to the critical FF’s of the circuit. The monitor outputs are used to define failure probability of the technique. For comparison the simulations are carried out under four different cases:

- V_{dd} reduction: as suggested in [10]
- Temperature reduction: as suggested in [10]
- Gate Oversizing: as suggested in [12]
- Time borrowing: as proposed in this paper

Fig.12 shows a comparison of the failure probability observed under the four considered cases. The figure shows that ReverseAge ensures 4.5 times more reliability as compared to temperature tuned proposed in [10] and 3.0 times lower failure probability than transistor sizing suggested by [12]. This means that the ReverseAge improves NBTI tolerance in the circuit by 3 times with respect to the state-of-the-art. This improvement proves the effectiveness of the proposed ReverseAge technique.

ReverseAge Overhead

Overheads of the ReverseAge includes area and power consumed by the monitor and tolerance circuit. The area
overhead is obtained in percent i.e. the area of the added monitor and tolerance circuit with respect to the original circuit. Fig. 13(a) shows percent the area overhead for multiple monitors in the circuit. Fig. 13 shows that the ReverseAge adds only 3.77% area overhead in comparison to the 8.7% overhead proposed in [12]. Similarly, the power overhead of ReverAge is only 1.8% and is very low in comparison to the power consumed by the NBTTI mitigating techniques proposed in [12,21].

VI. CONCLUSION

This paper presented ReverseAge, an online technique to tolerate NBTTI induced delays by utilizing design time margin. ReverseAge proposes an NBTTI monitoring scheme along with a technique to tolerate NBTTI induced delay. The monitor detects NBTTI induced delay and generate a timing error signal. Thereafter, the timing error of the stage is tolerated by borrowing time from the successive stages. Simulation result from benchmark circuit shows that ReverseAge ensures $3 \times$ reliability improvement at the cost of 3.77% area and 1.4% power overheads.

REFERENCES


