On-Line Prediction of NBTI-induced Aging Rates

Rafal Baranowski*, Farshad Firouzi†, Saman Kiamehr†, Chang Liu*, Mehdi Tahoori†, and Hans-Joachim Wunderlich*
* Institute of Computer Architecture and Computer Engineering, University of Stuttgart, Germany
† Karlsruhe Institute of Technology, Karlsruhe, Germany

***Abstract*—Nanoscale technologies are increasingly susceptible to aging processes such as Negative-Bias Temperature Instability (NBTI) which undermine the reliability of VLSI systems. Existing monitoring techniques can detect the violation of safety margins and hence make the prediction of an imminent failure possible. However, since such techniques can only detect measurable degradation effects which appear after a relatively long period of system operation, they are not well suited to early aging prediction and proactive aging alleviation.**

This work presents a novel method for the monitoring of NBTI-induced degradation rate in digital circuits. It enables the timely adoption of proper mitigation techniques that reduce the impact of aging. The developed method employs machine learning techniques to find a small set of so called Representative Critical Gates (RCG), the workload of which is correlated with the degradation of the entire circuit. The workload of RCGs is observed in hardware using so called workload monitors. The output of the workload monitors is evaluated on-line to predict system degradation experienced within a configurable (short) period of time, e.g. a fraction of a second. Experimental results show that the developed monitors predict the degradation rate with an average error of only 1.6% at 4.2% area overhead.

***Keywords*-Representative critical gates, workload monitoring, aging prediction, NBTI**

I. Introduction

As the scaling of technology nodes proceeds, Negative Bias Temperature Instability (NBTI) becomes a major threat to the reliability of VLSI devices [1]. NBTI consists in the oxide degradation of PMOS transistors that results in a gradual shift in the threshold voltage, which in turn causes an increased propagation delay. Eventually, NBTI stress may significantly increase the critical path delay and lead to timing violations.

Traditional approaches for NBTI monitoring measure *degradation effects*, i.e. provide an aggregated measure of the degradation that took place over a long period of time. They can be used to guide *reactive* techniques that manage the degradation effects, e.g. by frequency and voltage scaling [2] or adaptive body biasing [3]. Such monitoring techniques, however, are unable to track the *degradation rate*, i.e. the amount of stress caused by the currently running application. Degradation rate monitoring is crucial for the timely adoption of preventive techniques that alleviate aging, such as proactive frequency and voltage scaling [4] or dynamic cooling [5].

The goal of this work is to enable the monitoring of NBTI-induced *degradation rate* in digital circuits, i.e. to predict the increase of the critical path delay over a short period of time, long before any measurable degradation takes place. This kind of monitoring enables the adoption of novel *proactive* countermeasures that reduce the degradation rate, e.g. NBTI-aware scheduling, load balancing, frequency and voltage scaling, or guide the application of healing patterns [6].

In digital circuits, monitoring of the delay degradation rate is challenging for three reasons: (1) Direct aging rate estimation by consecutive measurements of the critical path delay is impractical over a short period of time as the delay increase is not measurable. (2) The degradation rate of a digital circuit depends on the degradation rate of each PMOS transistor on the current critical path, whereas the critical path may change over time. (3) The degradation rate of each PMOS transistor is a function of its state (or duty cycle) which in turn depends on the currently running application.

In this paper, we describe an innovative monitoring approach that combines workload monitoring with machine learning techniques. Our method is based on the monitoring of *representative critical gates* (RCG), the workload of which correlates with the degradation rate of the entire circuit. We provide an algorithm for finding a small set of RCGs, a method for the synthesis of RCG workload monitors, and an algorithm for on-line prediction of the delay degradation rate. Our experimental results show that the developed monitoring scheme predicts the degradation rate with an average error of only 1.6% at the expense of a modest 4.2% area overhead.

The rest of this paper is organized as follows: In the next section, we give the problem statement together with an overview of the developed monitoring scheme. Section III defines critical gates and provides the algorithm to select representative critical gates (RCGs). The synthesis of workload monitors for RCGs is described in Section IV. Section V deals with the construction of regression models that are used on-line to predict the degradation rate. The accuracy and overhead of the monitoring scheme is evaluated in Section VI.

II. Overview

A. Problem Statement

NBTI causes a gradual increase in gate delays and results in performance degradation of digital circuits. The aging rate at time t is defined as $\delta D(t)/\delta t$, where $D(t)$ is the length of the critical path at time t. The NBTI-induced aging rate is a function of many technology parameters, temperature, and the *duty cycle* of transistors, i.e. the ratio between transistor stress time to the total operating time.

NBTI is a long term phenomenon which impacts the circuit delay after a long time. To characterize the currently running application and guide aging alleviation techniques, it is sufficient to average the degradation rate over a period of several milliseconds to several minutes. Therefore, the goal of our work is to approximate the *average aging rate* over a given period of time T_0. At time $t = kT_0$, $k \in \mathbb{N}^+$, the average aging rate is calculated as $\Delta D/T_0$, where ΔD is the increase (degradation) of the circuit delay within the time interval $[t - T_0, t)$. The parameter T_0, i.e. the length of the time window over which the aging rate is averaged, must be configurable to suit different applications of aging rate monitoring.

B. Developed Monitoring Approach

The degradation rate monitoring scheme is presented in Fig. 1. The monitored circuit is augmented with a *workload*

978-3-9815370-4-8/DATE15/©2015 EDAA

monitor that observes a subset of the circuit's primary and pseudo-primary inputs, and a temperature sensor. Based on the state of the primary and pseudo-primary inputs, the workload monitor predicts the current stress of each representative critical gate (RCG) of the monitored circuit. Note that the RCGs are not monitored directly to limit the impact of monitoring on circuit performance. The outputs of the workload monitor are aggregated over a short period of time and used to predict the current degradation rate, i.e., the increase in the circuit delay over the recent period. A software component is responsible for the evaluation of the degradation rate based on the aggregated output from the workload monitor and the temperature sensor.

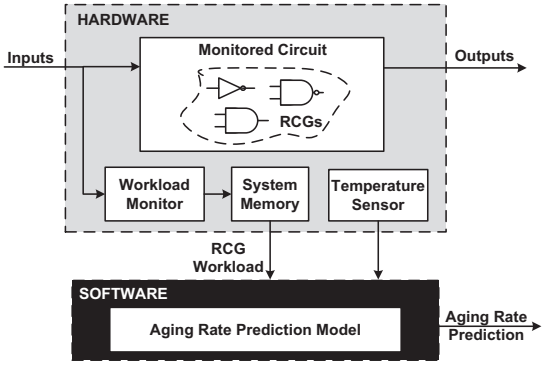

Fig. 1. Principle of the degradation rate monitoring

The monitoring method comprises design-time and on-line algorithms and activities, as shown in Fig. 2. At design-time, we characterize the target standard cell library w.r.t. NBTI aging and subject the circuit to an NBTI-aware Static Timing Analysis (STA). Based on STA results, we identify the Critical Gates (CG), i.e. gates that belong to critical paths or paths that may become critical due to aging. From the set of CGs, we select a small set of representative critical gates (RCGs), the delay degradation of which is correlated with the degradation of the entire circuit. Next, we synthesize workload monitors that predict the NBTI-relevant stress of RCGs. Finally, a regression-based aging rate prediction model is constructed.

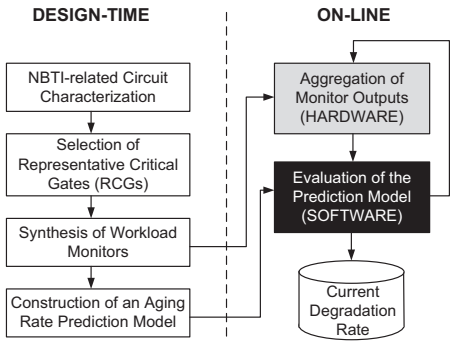

Fig. 2. Design-time algorithms and on-line activities for aging rate prediction

In a running system, the workload monitors continuously observe the stress experienced by the RCGs. Together with the current temperature, this information is periodically fed to the aging rate prediction model. The aging rate model is evaluated in software, ideally during the idle time of any available processing unit.

III. Representative Critical Gates (RCG)

A. Critical Gate Selection

We define the *slack of a gate* as the difference between the delay of the longest path through that gate and the critical path delay of the circuit. If the slack of a gate does not exceed a given threshold, the gate is called *critical gate* (CG). The set of CGs is found using an aging-aware timing analysis based on [7]. The threshold can be adjusted based on the delay degradation rate of the circuit.

We define the *workload of a critical gate* as the duty cycles of its constituent PMOS transistors. Circuit degradation rate is predicted by monitoring the workload of selected CGs, as discussed below.

B. RCG Selection Algorithm

Since monitoring of all the CGs in a circuit is infeasible, it is crucial to find a small set of representative critical gates (RCGs). The workload of RCGs must be highly correlated with the NBTI-induced delay degradation of the circuit. RCG selection problem can be formulated as follows: for a given set of CGs V, identify a minimal subset $F \subseteq V$ such that the circuit delay degradation rate can be predicted with sufficient accuracy only from the workload of gates in F and the chip temperature. Solving this problem requires all possible circuit applications and aging histories be considered, which is infeasible. Instead, we select the set of RCGs by evaluating the correlation between the delays of RCGs and the circuit delay.

The set of RCGs is found using the *wrapper* method for feature selection [8]. The training set consists of sets of degraded CG delays and the corresponding degraded circuit delay obtained for random applications. The following merit function is used to score each subset $S \subseteq V$ consisting of k features: $k\overline{r_{cf}}/\sqrt{k+k(k-1)\overline{r_{ff}}}$, where $\overline{r_{ff}}$ is the correlation between features, and $\overline{r_{cf}}$ is the average correlation between features and circuit delay. The correlations $\overline{r_{ff}}$ and $\overline{r_{cf}}$ are calculated as *minimum description length* [8]: $E(A) + E(B|A)$, where E is entropy while A and B are either feature or delay degradation.

IV. Workload Monitoring

A *workload monitor* is a combinational circuit that observes a subset of the circuit's primary inputs and pseudo-primary inputs (register outputs) and generates one output per PMOS transistor of each RCG. An output of the monitor is "1" when the corresponding transistor experiences NBTI stress ($\mathtt{U_{GS}} < 0$), and it is "0" otherwise. Since the monitor observes only primary and pseudo-primary inputs, it has minimal impact on the timing of the monitored circuit.

The problem of monitor synthesis is defined as follows: Given a circuit with n primary inputs and m pseudo-primary inputs, a set of RCGs, and target monitoring accuracy A_T expressed in percent, construct a workload monitor with minimal area overhead such that each monitor output provides the correct response for at least $A_T \cdot 2^{n+m}$ input patterns.

In principle, workload monitors can be synthesized using the technique presented in [9]. Our experimental results show, however, that this approach results in prohibitive area overhead

that often exceeds 100% for the usual number of RCGs required for accurate aging rate prediction. In the following, we describe a novel heuristic method for the synthesis of workload monitors with affordable area overhead: Initially, we construct an *exact monitor* that provides correct response for all input patterns. Next, we build *approximate monitors* with reduced size by iterative reduction of the exact monitor.

The exact monitor is simply a copy of the monitored circuit in which (1) all gates that do not belong to the transitive fan-in cone of any RCG are removed, (2) additional gates are added to generate outputs that are "1" whenever their corresponding PMOS transistors experience NBTI stress ($\mathtt{U_{GS}} < 0$).

An approximate monitor is constructed by iterative removal of gates from the exact monitor at the cost of reduced accuracy. A gate is removed if its output controllability and observability is low enough to guarantee the target monitoring accuracy A_T. The removal of gates is modeled by the injection of *stuck-at faults* in the monitor. For each injected fault, we evaluate the resulting accuracy, as well as the area reduction due to fault injection and constant propagation. In every iteration, we select a fault that does not violate the target monitoring accuracy A_T and results in the lowest size of the monitor.

Formally, the approximate monitors are synthesized as follows: Let M be the exact monitor with an output set $\{o_1, o_2, o_3, \ldots, o_G\}$. Iterate:

1) Generate a structurally collapsed stuck-at fault set $\{f_1, f_2, f_3, \ldots, f_K\}$ for M.
2) For each fault f_i, create a structural copy of M, denoted M_i, with injected fault f_i.
3) Using Monte Carlo simulation, evaluate the accuracy of each monitor output o_j of M_i, denoted $a_i(o_j)$, i.e. the ratio of input patterns for which the output o_j in M_i matches the corresponding output in the *exact monitor*.
4) For each faulty monitor M_i, calculate $A_i = \mathtt{MIN}_{j=1}^{G}\ a_i(o_j)$.
5) Terminate if $\mathtt{MAX}_{i=1}^{K}\ A_i < A_T$.
6) For each M_i with $A_i \geq A_T$, perform fault/constant propagation, remove the gates with constant outputs, and simplify the gates with constant inputs.
7) Approximate the size of each M_i, denoted S_i, as the total area of standard cells in M_i.
8) Choose monitor M_t, $1 \leq t \leq K$, such that $A_t \geq A_T$ and $S_t = \mathtt{MIN}_{i=1}^{K} S_i$; assign $M := M_t$, repeat from step 1.

V. Prediction of the Degradation Rate

A. Off-Line Construction of the Aging Rate Model

For on-line prediction, it is necessary to model the relationship between the output of workload monitors and the delay degradation rate of the circuit. The aging rate model is a linear regression of the following form: $\Delta D = W \cdot \beta + \epsilon$, where ΔD is the increase in circuit delay, W is a vector of average RCG workloads from the monitor, while the vector β and the scalar ϵ are regression coefficients. To fit the model, we use the *ordinary least squares* (OLS) technique.

B. On-Line Model Evaluation

To evaluate the aging rate prediction model, the duty cycles of PMOS transistors in RCGs must be averaged over the time period T_0. To this end, each output of the workload monitor is aggregated in hardware. In the simplest case, one counter per monitor output is used to count up the number of clock cycles in which the output is "1", i.e. cycles in which the corresponding PMOS transistor experiences NBTI stress.

To reduce the monitoring cost, an existing memory system is reused and systematic sampling methods are applied: Just a single counter with multiplexed input is used regardless of how many RCGs are monitored. The counter is attached to each monitor output for the time period of T_0/G, where G is the total number of monitor outputs. The minimal counter length is hence $L := \lceil log_2 \frac{T_0}{G \cdot T_{clk}} \rceil$, where T_{clk} is the minimal period of the system clock. The content of the counter is sent to the main system memory with a period of T_0/G using Direct Memory Access (DMA). The memory demand is $G \cdot L$ bits. For instance, assuming 30 monitor outputs, an evaluation period of 1 s, and 1 GHz clock, only a single 25-bit counter and 94 bytes of memory are required.

Since the aging rate prediction is executed relatively seldom, e.g. with a period of several hundred milliseconds to minutes, it is performed in software during the idle time of any currently available processing unit. The processing unit reads the average chip temperature as well as the aggregated output of the workload monitors from the system memory. These data are then used to evaluate the aging rate prediction model. The only requirement on the processing unit is that it supports efficient addition and multiplication of fixed or floating point numbers.

VI. Evaluation

A. Experimental Setup

We evaluate the on-line prediction technique on a set of ISCAS'89 benchmark circuits. We define an application of a benchmark circuit as a set of primary input signal probabilities. The Nangate 45nm open cell library [10] is used to synthesize the benchmark circuits and monitors. We assume that the benchmarks are part of a system with a temperature sensor, a memory system with DMA and free capacity of several KB, and a processing unit that can be reused in the idle time to evaluate the prediction model. We exploit the *reaction-diffusion* (R-D) NBTI model proposed in [11]. The worst case NBTI-induced delay degradation is assumed to be 10% over 3 years for a simple inverter and the parameters of the NBTI model are set accordingly. Due to limited space, we provide results for a constant system temperature.

B. Impact of Application on Aging Rate

To study the effect of different applications on the aging rate, 5000 random sets of primary input signal probabilities are considered. Fig. 3 shows the range by which the degradation rate differs across the applications for each benchmark circuit. The range is calculated as: $[\max(\Delta D) - \min(\Delta D)]/\min(\Delta D)$, where $\max(\Delta D)$ and $\min(\Delta D)$ are respectively the maximum and minimum NBTI-induced delay degradation over all applications. Since the effect of the application on the amount of NBTI-induced delay degradation can be as large as 40%, it is crucial to monitor the circuit's workload for an accurate aging rate prediction.

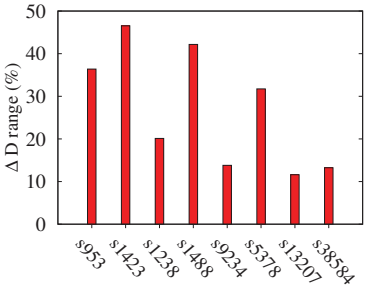

Fig. 3. Normalized range of NBTI-induced delay degradation across different applications

C. Validation Experiments

To evaluate the accuracy of the developed method, 4500 of the application sets are used as a training set for linear regression, and 500 sets are used for validation. For each application, the accurate NBTI-induced delay degradation ΔD_{calc} is computed based on the NBTI-aware timing analysis [7]. The corresponding on-line estimation ΔD_{est} is obtained by feeding the average outputs of the workload monitor to the aging rate prediction model. The prediction accuracy is calculated as *normalized root-mean-square error* (NRMSE):

$$NRMSE = \frac{\sqrt{\sum_{i=1}^{n}(\Delta D_{est_i} - \Delta D_{calc_i})^2}}{\sqrt{n}(\Delta D_{calc_{max}} - \Delta D_{calc_{min}})}, \qquad (1)$$

where n is the number of applications (in this case 500).

D. Accuracy of Aging Prediction

Table I shows the number of CGs for all benchmark circuits assuming a CG slack threshold of 5%. Simulation results show that 5% slack threshold gives a good trade-off between the number of CGs and prediction accuracy. As described in Section III-B, machine-learning is used to find a set of RCGs to decrease the monitoring overhead. The column 4 of Table I shows the number of RCGs, which is less than the number of CGs by a factor of four on average.

We evaluate the prediction accuracy with two sets of workload monitors: the first one can monitor the workload with 100% accuracy and the other with 80% accuracy (see Section IV). As shown in Fig. 4, the average NRMSE of the proposed technique is around 1.8% and 3.6% for 100% and 80% monitors, respectively. These results correspond to a mean error of 0.8% and 1.6% for 100% and 80% monitors, respectively. While the inaccuracy of 80% monitors is higher, their area overhead is less as discussed below.

E. Hardware Overhead

The area overhead of the workload monitors is presented in Table I: the size of the benchmark circuits is given in the fifth column, while the two following columns show the monitoring overhead for monitors with the target accuracy of 100% and 80% (cf. Section IV). The area overhead of the 80% monitors, which were shown to provide sufficient accuracy, is below 2.4% for circuits with more than 1000 gates. The monitoring overhead scales well with the circuit size: for the two largest benchmarks, the overhead is below 0.3%.

TABLE I: AREA OVERHEAD OF THE DEVELOPED MONITORING SCHEME

Circuit	#Gates	#CGs	#RCGs	Area [μm^2]	Monitoring overhead	
					$A_T = 100\%$	$A_T = 80\%$
s953	683	26	18	479.86	24.39%	8.31%
s1423	824	116	44	808.64	22.13%	10.23%
s1238	881	47	22	546.10	56.06%	7.54%
s1488	902	27	14	571.37	13.73%	3.49%
s9234	1725	92	15	2388.68	15.93%	1.61%
s5378	2926	50	16	1934.88	6.61%	2.36%
s13207	4074	162	14	5159.87	15.10%	0.20%
s38584	18142	59	21	16861.74	0.48%	0.30%

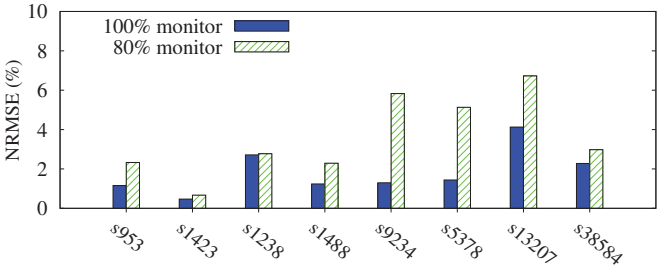

Fig. 4. Degradation rate prediction accuracy. NRMSE: normalized root-mean-square error

F. Runtime

As mentioned in Section V, the prediction technique consists of the off-line monitor synthesis and construction of the aging rate model, and the on-line model evaluation. The runtime of the off-line part is only few hours for the largest circuit. The on-line prediction time is negligible as it involves simple software vector multiplication that is executed relatively seldom.

VII. CONCLUSION

Since the NBTI effect in advanced technology nodes strongly depends on the system application and workload, on-line prediction of the degradation rate is crucial to enable effective aging alleviation. We present a novel method for aging rate prediction which is based on workload monitoring and machine learning techniques. The monitoring technique enables on-line prediction of the degradation rate caused by the currently running application. Experimental results show that this method delivers sufficient accuracy at an affordable area overhead which decreases with the size of the monitored circuit.

ACKNOWLEDGMENT

This work was supported by the German Research Foundation (DFG) under grants TA 782/9-1 and WU 245/13-1 (RM-BIST).

REFERENCES

[1] M. Alam *et al.*, "A Comprehensive Model of PMOS NBTI Degradation," *Microelectronics Reliability*, vol. 45, no. 1, pp. 71–81, 2005.
[2] M. Nakai *et al.*, "Dynamic Voltage and Frequency Management for a Low-Power Embedded Microprocessor," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 1, pp. 28–35, 2005.
[3] S. Kumar *et al.*, "Adaptive Techniques for Overcoming Performance Degradation due to Aging in Digital Circuits," in *Proc. IEEE Asia and South Pacific Design Automation Conf. (ASP-DAC)*, 2009, pp. 284–289.
[4] C. Zhuo *et al.*, "Process Variation and Temperature-Aware Reliability Management," in *Proc. Design, Automation and Test in Europe (DATE)*, 2010, pp. 580–585.
[5] E. Mintarno *et al.*, "Optimized Self-Tuning for Circuit Aging," in *Proc. Design, Automation and Test in Europe (DATE)*, 2010, pp. 586–591.
[6] F. Firouzi *et al.*, "NBTI Mitigation by Optimized NOP Assignment and Insertion," in *Proc. Design, Automation Test in Europe (DATE)*, March 2012, pp. 218–223.
[7] F. Firouzi *et al.*, "Incorporating the Impacts of Workload-Dependent Runtime Variations into Timing Analysis," in *Proc. Design, Automation and Test in Europe (DATE)*, 2013, pp. 1022–1025.
[8] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.
[9] R. Baranowski *et al.*, "Synthesis of Workload Monitors for On-Line Stress Prediction," in *Proc. IEEE Intl. Symp. on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT)*, 2013, pp. 137–142.
[10] Nangate 45nm Open Cell Library v1.3, http://www.nangate.com.
[11] S. Bhardwaj *et al.*, "Predictive Modeling of the NBTI Effect for Reliable Design," in *Proc. IEEE Custom Integrated Circuits Conf. (CICC)*, 2006, pp. 189–192.