ARTICLE IN PRESS

J. Parallel Distrib. Comput. ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

J. Parallel Distrib. Comput.

journal homepage: www.elsevier.com/locate/jpdc

Critical transistors nexus based circuit-level aging assessment and prediction

N. Cucu Laurenciu *, S.D. Cotofana

Computer Engineering Laboratory, Delft University of Technology, Delft 2628CD, The Netherlands

HIGHLIGHTS

- The paper proposes two methods to assess a circuit reliability (end-of-life).
- The Markovian framework accounts for process, environmental and temporal variations.
- The topology dependent framework uses the end-of-life of the critical transistors.
- The frameworks are validated and their estimation accuracies are assessed.
- Area (number of critical transistors) vs. accuracy trade-offs are investigated.

ARTICLE INFO

Article history:
Received 25 February 2013
Received in revised form
29 July 2013
Accepted 3 August 2013
Available online xxxx

Keywords:
Design for reliability
FEOL reliability
Transistor aging
Markovian aging model
Circuit aging

ABSTRACT

Accurate age modeling, and fast, yet robust reliability sign-off emerged as mandatory constraints in Integrated Circuits (ICs) design for advanced process technology nodes. In this paper we introduce a novel method to assess and predict the circuit reliability at design time as well as at run-time. The main goal of our proposal is to allow for: (i) design time reliability optimization; (ii) fine tuning of the run-time reliability assessment infrastructure, and (iii) run-time aging assessment. To this end, we propose to select a minimum-size kernel of critical transistors and based on them to assess and predict an IC End-Of-Life (EOL) via two methods: (i) as the sum of the critical transistors end-of-life values, weighted by fixed topology-dependent coefficients, and (ii) by a Markovian framework applied to the critical transistors, which takes into account the joint effects of process, environmental, and temporal variations. The former model exploits the aging dependence on the circuit topology to enable fast run-time reliability assessment with minimum aging sensors requirements. By allowing the performance boundary to vary in time such that both remnant and nonremnant variations are encompassed, and imposing a Markovian evolution, the probabilistic model can be better fitted to various real conditions, thus enabling at design-time appropriate guardbands selection and effective aging mitigation/compensation techniques. The proposed framework has been validated for different stress conditions, under process variations and aging effects, for the ISCAS-85 c499 circuit, in PTM 45 nm technology. From the total of 1526 transistors, we obtained a kernel of 15 critical transistors, for which the set of topology dependent weights were derived. Our simulation results for 15 critical transistors kernel indicate a small approximation error (i.e., mean smaller than 15% and standard deviation smaller than 6%) for the considered circuit estimated end-of-life (EOL), when comparing to the end-of-life values obtained from Cadence simulation, which quantitatively confirm the accuracy of the IC lifetime evaluation. Moreover, as the number of critical transistors determines the area overhead, we also investigated the implications of reducing their number on the reliability assessment accuracy. When only 5 transistors are included into the critical set instead of 15, which results in a 66% area overhead reduction, the EOL estimation accuracy diminished with 18%. This indicates that area vs. accuracy trade-offs are possible, while maintaining the aging prediction accuracy within reasonable bounds.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

Wear out mechanisms [26], further aggravated by the aggressive CMOS scaling adopted for performance improvement, have emerged as major reliability concerns for deep sub-micron devices [14,13,3]. The time dependent drift of critical physical and electrical transistor parameters, due to manufacturing and environmental induced variations, as well as run-time aging effects, degrades the performance and eventually produces device failure. These considerations, in addition with the high pressure in achieving short time-to-market figures extended the need for reliability

* Corresponding author.
E-mail address: N.CucuLaurenciu@tudelft.nl (N. Cucu Laurenciu).

0743-7315/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jpdc.2013.08.006

Please cite this article in press as: N. Cucu Laurenciu, S.D. Cotofana, Critical transistors nexus based circuit-level aging assessment and prediction, J. Parallel Distrib. Comput. (2013), http://dx.doi.org/10.1016/j.jpdc.2013.08.006

analysis also in early development stages, e.g., at the design time [8,27,33].

Most of past approaches that address the circuit-level reliability analysis mainly focus on either temporal variations—caused by aging mechanism such as Negative Bias Temperature Instability (NBTI), Hot Carriers Injection (HCI), time dependent dielectric breakdown, electromigration, thermal cycling—[15,2], or on process variations [17,16], without considering the interactions between them. Only recently, studies considering joint effects have been reported in the literature. In the digital domain, aging-aware Statistical Timing Analysis (STA) schemes that rely on analytical expressions of circuit performance features (e.g., propagation delay, signal slope) as a function of process/wearout degradation parameters have been proposed. In [31], based on device parameters statistical spread shifts, the circuit delay fall-out is obtained as an indicator of process variations and NBTI aging effect. In [19] a Statistical Static Timing Analysis method (SSTA) is proposed in order to characterize the circuit delay distribution under process variations and NBTI effects. [34] introduces a statistical age prediction framework for a circuit path under process variations and temporal stress. In [24] an analytical model suitable for circuit level that captures both short term NBTI and process variations effects is developed and used to quantify their impact on the circuit nominal degradation. In [32], the authors introduce the concept of virtual age that reflects the circuit cumulative aging evolution and propose a real time circuit time-to-failure prediction framework.

We note that previous approaches towards aging models are deterministic. However, due to the very nature of the aging inducing phenomena we believe that a more appropriate, but also more complex approach should be a full probabilistic model. In this way the age could be regarded not only as a function of the instantaneous value at time t of a degradation parameter X, for example, but also of its history (from $t = 0$ to the time moment t at which we want to compute the age):

$$A = A(t, x_1, x_2, \ldots, x_n), \tag{1}$$

where $x_1, x_2, \ldots, x_n$ are stochastic processes which enter in the expression of A by their particular realizations. As a consequence, A is also a stochastic process whose characteristics (e.g., probabilities, moments) have to be obtained from the properties of $x_1, x_2, \ldots, x_n$. This is a very general formulation and for a workable model, obviously, we have to impose particular restrictions.

The simplest and roughest simplification of this dependency is to express the age solely as a function of the parameter values at time moment t:

$$A = A(x_1(t), x_2(t), \ldots, x_n(t)). \tag{2}$$

This brings us back to the point of view adopted in previous deterministic approaches, thus we do not follow this avenue.

Another simplification can be made based on the fact that we do not need all the values between 0 and t but only the values in a finite number of moments. In fact, we can further assume that only the value at the current time moment, (denoted in the sequel by $x_i(t_k)$) and the one at the previous sampling moment (denoted from now on by $x_i(t_{k-1})$) are required. In the general case $x_i(t_k)$ and $x_i(t_{k-1})$ are not independent random variables, but correlated and passing from one to the other could be governed by probabilistic laws. The processes x_i could be Markovian processes and this character could be transferred to A. Moreover, the processes $x_1, x_2, \ldots, x_n$ could be correlated. In this case, if we describe (via a change of variables) A as a function of other processes $X_1, X_2, \ldots, X_n$ obtained from $x_1, x_2, \ldots, x_n$ by a linear transform of Karhunen–Loeve (KL) type [7], the process A can be approximated by making use of a small number of variables. In this manner, one can obtain a correct description of A by, e.g., a function of 4 variables $X_1(k),\ X_2(k),\ X_1(k-1),\ X_2(k-1)$. In view of the above, the following remark is in order: a Markovian model fitted to the age problem must have the transition probabilities not only time dependent but also dependent of the new states. Our approach introduces a Markovian model fitted to the circuit-level aging problem. Furthermore, instead of considering a fixed performance boundary, we allow it to vary in time. In this way we obtain a more flexible model, which takes into consideration that depending on stress duration, the effects on the circuit statistical parameters could be remnant or nonremnant. As a result, guardbands selection and appropriate aging mitigation/compensation techniques, better fitted to real working conditions are enabled.

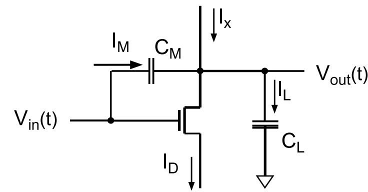


Fig. 1. Circuit schematic for transistor age assessment.

In view of the previous discussion, this paper proposes: (i) the selection of a minimum size kernel of critical transistors based on which the circuit end-of-life can be estimated; (ii) a run-time aging framework that estimates the circuit end-of-life as the sum of critical transistors end-of-life values weighted by fixed, topology dependent coefficients; and (iii) a Markovian aging framework that is capable of assessing and predicting the circuit performance degradation and lifetime.

The proposed critical transistors kernel based aging assessment and prediction framework is validated by means of simulation. The simulation is performed in Cadence Relxpert and Spectre, and Synopsys Pathmill, using as test circuit the ISCAS-85 c499, implemented in PTM 45 nm technology. Exposing the circuit to several stress profiles, from a total of 1526 transistors, a kernel of 15 critical transistors and their corresponding topology dependent weights were obtained. When subjecting the ISCAS-85 c499 circuit to new sets of stress profiles and comparing the circuit end-of-life estimated with the proposed framework against the results from Cadence and Pathmill, relatively small values of the approximation error (i.e., mean smaller 10% and standard deviation smaller than 6%) are obtained, which quantitatively validate and confirm the lifetime prediction accuracy of proposed framework. Moreover, as the number of critical transistors determines the area overhead, we also investigated the implications of reducing their number on the reliability assessment accuracy. When only 5 transistors are included into the critical set instead of 15, which results in a 66% area overhead reduction, the EOL estimation accuracy diminished with 18%. This indicates that area vs. accuracy tradeoffs are possible, while maintaining the aging prediction accuracy within reasonable bounds.

The rest of the paper is organized as follows: Section 2 briefly describes the transistor-level aging assessment framework. Section 3 extends the transistor-level aging framework to the circuit level, which is introduced in Section 4. The simulation methodology and the obtained results are presented in Section 5. The paper is concluded in Section 6 with some final remarks.

2. Transistor-level aging framework

Fig. 1 presents the circuit we utilize for transistor age assessment [9,10]. In the following, we note with P_{in} the slope of the gate voltage V_{in}, P_{out} the slope of the output voltage, and P_x the slope of the surrounding current contribution I_x.

For proper transistor lifetime characterization, one should take into account not only the intrinsic self-degradation, but also

the surrounding circuit topology influence on the transistor in question (i.e., the influence of adjacent degraded transistors on the transistor under study). For instance, we propose to account for the influence of adjacent transistors, by means of:

- the variation of gate voltage slope $\Delta dV_{in}/dt$, which captures the impact of driver transistors aging,
- the variation of I_x current slope $\Delta dI_x/dt$, which captures the impact of the aging of transistors connected to the source terminal.

For the purpose of illustration, we consider as performance parameter the slope P_{out} of the output voltage. The modification of the voltage slope from device input to its output, namely: $P_\Delta = P_{in} - P_{out}$, measures the influence of the device in the signal degradation. We express P_Δ as being composed out of two terms:

$$P_\Delta = P_{\Delta a} + P_{\Delta 0}, \tag{3}$$

where $P_{\Delta 0}$ accounts for the inherent and initial slope degradation; and $P_{\Delta a}$ is the degradation part which increases with the device age. In principle, the term $P_{\Delta a}$ accounts for all factors which negatively impact the device performance, that are: (i) intrinsic factors—the drift of own degradation parameters $\mathbf{X}$ (e.g., Vth), and (ii) extrinsic factors—the variation of P_{in} and P_x slopes. More formally, $P_{\Delta a}$ can be expressed as a functional as follows:

$$P_{\Delta a} = f(\mathbf{X}(.),\ P_{in}(.),\ P_x(.)). \tag{4}$$

In view of the above, the age can be defined through the time integral $\int_0^t dP_{\Delta a}dt$, where $dP_{\Delta a} = dA$ are the time decrements of the slope. Once the aging increment is computed, one can proceed with the derivation of the aging rate and age expressions, i.e., the aging rate is derived by taking the ratio between the aging and time increments $\left(A_{rate} = \frac{dP_{\Delta a}}{dt}\right)$ and the age is given by integrating the aging increment over the interval $[0,\ t]$.

In the analysis of a single transistor one can consider P_{in} as being constant (i.e., the input signal is always not degraded) and in general we normalize the age such that $A = k \cdot (P_{in} - P_{out})$ arrives at the value 1 when P_{out} arrives at 0.9 of its initial value P_{out0}, for a given standard value of P_{in}. In consequence, for estimating the age of a transistor in real operating conditions we have to compute the value:

$$A = k \cdot \int_0^t (dP_{in} - dP_{out})dt = k \cdot \int_0^t A_{rate}dt. \tag{5}$$

In the next section, making use of the transistor-level aging model, we propose a modality to find the location and number of monitored transistors required to determine the overall circuit performance degradation.

3. Circuit performance degradation

The time-dependent wearout, i.e., aging, affecting a circuit transistors, is reflected at the circuit level as degradation of its performance parameters, such as the increase of the circuit propagation delay. Eventually, the age-induced circuit propagation delay degradation, can exceed the maximum circuit clock period and as a consequence, wrong values may be sampled and hence circuit erroneous functioning induced as the circuit reaches its end-of-life.

In order to estimate a circuit end-of-life, we propose to express: (i) a circuit End-Of-Life (EOL) as the minimum end-of-life of its propagation paths, and (ii) a propagation path end-of-life as the sum of the end-of-life values of all its comprising transistors, weighted by topology-dependent coefficients. Let us consider a circuit and denote by M the number of its propagation paths. The circuit end-of-life can then be expressed as follows:

$$EOL_{circuit} = \min_j (EOL_{path\,j}), \tag{6}$$

$$EOL_{path\,j} = \sum_{i=1}^{N_j} w_{ij} \cdot EOL_i, \tag{7}$$

where $j = 1, \ldots, M$, N_j is the number of transistors contained by path j, w_{ij} are topology dependent coefficients, EOL_i represents the end-of-life of transistor i, and $EOL_{path\,j}$ represents the end-of-life of path j.

However, this approach is not feasible, as embedded wear-out sensors are expensive in terms of silicon area and real life circuits may encompass thousands of paths and millions of transistors. A reduction of the number of wear-out measurement sites is thus required for tractability purposes of circuit aging derivation. To this extent, the following model simplifications are made: (i) we reduce the number of paths to a set of critical ones, and (ii) we reduce the numbers of transistors to a kernel set. The model thus becomes:

$$EOL_{circuit} = \min_j (EOL_{path\,j}) \tag{8}$$

$$EOL_{path\,j} = \sum_{i=1}^{N_{reduced}} w_{ij} \cdot EOL_i, \tag{9}$$

where $j = 1, \ldots, M_{reduced}$, $M_{reduced}$ is the number of paths, and $N_{reduced}$ is the number of critical transistors. The transistors end-of-life values entering the above equations can be obtained for instance by utilizing the transistor level aging model proposed in [9]. In the sequel, we present the reduction criteria and the critical paths and critical transistors selection methodologies.

As far as the paths are concerned, we employ as reduction criterion, the path criticality in the circuit from the timing point of view. If the aging-induced degradation of a certain path P_1 is larger than that of the initial (unaged, at time 0) critical path P_0 (which determines the clock period), then the circuit timing constraints are violated, and P_1 becomes the circuit new critical path. Therefore, in order to assess the circuit reliability profile, we consider as critical paths the ones that could violate the timing constraints when their comprising transistors are subjected to wear-out induced degradation. By following this principle, the aging of the critical paths can be determined at design-time by performing aging-aware statical timing analysis [23].

As concerns the kernel of critical transistors, we note that for a critical path, only a small percentage of its transistors could potentially cause significant circuit performance degradation due to their aging. As a consequence, a critical path end-of-life can be estimated from a reduced subset of all its comprising transistors, i.e., the path's critical transistors. Thus, the kernel set can be formed as the reunion of the critical transistors for each critical path.

Even though a circuit path may comprise a plethora of transistors, some of them may be weakly correlated with the end-of-life of the critical paths, while others may be redundant in the estimation if their aging is being highly correlated with the aging of other transistors. This suggests the selection of a reduced, common kernel of critical transistors to be utilized for estimating the end-of-life of all the critical paths, as a more appropriate approach. More precisely, we are not interested in selecting the critical transistors that have aged the most, but in selecting the ones that are useful from a prediction point of view, e.g., the redundant but relevant – statistically dependent with the end-of-life of the critical paths – transistors can be excluded from the kernel of critical transistors. In view of the above, we propose to further reduce the cardinality of the critical transistor kernel, by estimating each critical path end-of-life from the same, common subset of critical transistors, regardless of their appartenance to a particular critical path. That is, instead of using a separate subset of transistors for each path, all of them belonging to the path whose end-of-life is being estimated,

Please cite this article in press as: N. Cucu Laurenciu, S.D. Cotofana, Critical transistors nexus based circuit-level aging assessment and prediction, J. Parallel Distrib. Comput. (2013), http://dx.doi.org/10.1016/j.jpdc.2013.08.006

we use a common kernel of transistors, not all belonging to the critical path whose end-of-life is being estimated. Thus we select the most relevant critical transistors, specifically the ones with the biggest impact on the circuit aging.

The task of selecting the critical transistor kernel reduces to a multi-response regression problem (i.e., estimating multiple response variables, i.e., paths end-of-lives, using a reduced, common kernel of input variables, i.e., transistor end-of-lives). Besides the benefits of critical transistors kernel cardinality reduction, using a single, unified model to estimate all the responses simultaneously exhibits also increased computational efficiency, and better prediction accuracy [4,1], when compared to building a separate model for each response variable.

The problem of selecting the critical transistors kernel, can be formalized as follows: Suppose we have n end-of-life measurements of the p critical paths and of the m transistors encompassed by the p paths. Let the response variables be denoted by a $n \times p$ matrix $\mathbf{Y} = [\mathbf{y_1} \dots \mathbf{y_p}]$, and the input variables by a $n \times m$ matrix $\mathbf{X} = [\mathbf{x_1} \dots \mathbf{x_m}]$. A linear model of the form:

$$\widehat{\mathbf{Y}} = \mathbf{X} \cdot \mathbf{W} \tag{10}$$

is employed for estimating the responses matrix $\mathbf{Y}$, where $\mathbf{W}$ denotes the unknown $m \times p$ regression coefficients matrix desired to have a minimal number q of non-zero rows. Hence q denotes the cardinality of the smallest subset of input variables used to synthesize all response variables. Matrix $\widehat{Y}$ consists of the end-of-life of the critical paths, for the n measurements; matrix $\mathbf{X}$ consists of the end-of-life of the critical transistors, and $\mathbf{W}$ contains the topology dependent weights.

The problem of selecting the kernel of critical transistors and determining the corresponding topology dependent coefficients can be formally stated as follows:

$$\min_{\mathbf{W}} \left\| \mathbf{W}^T \right\|_{l_0} \qquad s.t.\ \frac{1}{2} \left\| \mathbf{Y} - \widehat{\mathbf{Y}} \right\|_F^2 \le \epsilon, \tag{11}$$

where $\|\cdot\|_F$ is the Frobenius norm, that is $\|\mathbf{B}\|_F^2 = \sum_{i,j} b_{ij}^2$, and $\left\| \mathbf{W}^T \right\|_{l_0}$ is the l_0 norm of $\mathbf{W^T}$, defined as the cardinality of the set $\left\{ i \in \{1 \dots p\} :\ w_{i,k} \neq 0 \text{ for some } k \right\}$.

We note herein that a regression coefficient w_{ij} can be regarded as the importance the i-th input variable has on the j-th response. This optimization problem translates into minimizing the number of non-zero rows of the regression coefficients matrix $\mathbf{W}$, while keeping the estimation error below a certain bound—in our case, the error tolerance being a function of the circuit timing constraints. Since the norm $\left\| \mathbf{W}^T \right\|_{l_0}$ is a discrete valued function, it yields to a NP-hard problem in terms of computational complexity. The computational intractability can be addressed in two ways: either by using suboptimal algorithms, or by relaxing the problem as for instance via the replacement of the l_0 norm with a convex mixed-norm $l_{p,q}$, defined as:

$$\|\mathbf{B}\|_{l_{p,q}} = \sum_i \left\| b_{i,.} \right\|_q^p, \quad \text{where } \left| b_{i,.} \right|_q = \left(\sum_j \left| b_{i,j} \right|^q \right)^{1/q}, \tag{12}$$

among which the most practical instances are $l_{1,q}$ norms with $q \in \{1, 2, \infty\}$ [35,29,18]. In our case, we use the $l_{1,2}$ norm to quantify the importance of an input variable in synthesizing the response variables. We refer the reader to [25] for the algorithmic details concerning the $l_{1,2}$ optimization problem, and to [30] for the $l_{1,\infty}$ optimization problem.

At this point we have determined the set of critical paths, the set of critical transistors to be monitored by aging sensors, and their topology dependent coefficients w_{ij}. A circuit end-of-life can now be estimated at run-time, as the sum of the critical transistors end-of-life values (obtained from the aging sensors), weighted by the fixed topology dependent coefficients w_{ij} determined at design-time with the previously presented methodology.

If an increased accuracy of estimating the circuit end-of-life is required, a probabilistic aging model that takes into account the history of aging is better suited. This is the case of the Markovian model presented in the next section, which can estimate the circuit end-of-life based on the kernel of critical transistors obtained according to the above methodology.

4. Circuit level aging model

In the framework proposed in [10], we define the age of a circuit based on the kernel of its critical transistors, as a function of many parameters which can be divided into three main categories: (i) $\mathbf{d}$, design parameters (e.g., the channel width W), which are subject to optimizations; (ii) $\mathbf{s}$, statistical parameters (e.g., the threshold voltage Vth) that fluctuate during to manufacturing process but also evolve in time depending on the dynamic operating conditions—their random behavior can only be described in probabilistic terms as random processes; and (iii) $\mathbf{r}$, range parameters (e.g., the temperature T, the supply voltage V_{DD}) whose variations are handled by specifying the range of values that can be attained.

In the following, the relation between the degradation parameters X_i and the performance parameter P_{out} will be given by a function f:

$$f : \mathbb{R}^n \to \mathbb{R}; \qquad f(\mathbf{X}) = P_{out}, \quad \text{where } \mathbf{X} = \{X_i\}_{i=1\dots n}.$$

During the lifetime of a device, its performance has to be better than an imposed value, which in our case means:

$$P_{out} > P_{out\,min}. \tag{13}$$

As P_{out} is time dependent (more precisely decreases with as time is passing) through various parameters among which, some are random processes, the lifetime of the device can be expressed in probabilistic terms as:

$$R(t) = Prob\left\{P_{out}(t) > P_{out\,min}\right\}. \tag{14}$$

The device end of life is thus given by the value of t for which $P_{out}(t) = P_{out\,min}$.

Further, we adopt the usual method to achieve tractability of our problem, namely, in the case of more than one scalar statistical parameter, we apply on each of these parameters (with the restriction of having unimodal distributions) appropriate transforms to convert them into normal distributed random variables [12], while maintaining the correlation among each pair. In consequence, the statistical parameters become a normal distributed vector. In this way, we are able to compute the worst-case distance d_w, defined in [11] as the Mahalanobis distance between the mean point $\mathbf{s}_0$ and the worst case point $\mathbf{s}_w$ (i.e., the point belonging to the set of all parameters that violate a specification and is closest to the mean vector $\mathbf{s}_0$):

$$d_w^2 = (\mathbf{s}_w - \mathbf{s}_0)^T \cdot \mathbf{C}^{-1} \cdot (\mathbf{s}_w - \mathbf{s}_0), \tag{15}$$

where, as said, $\mathbf{s}$ denote the vector of statistical parameters after transforming them into Gaussian variables; $\mathbf{s_0}$ is its mean vector, and $\mathbf{C}$ is its covariance matrix, all at the same time moment. The worst case distance d_w, obtained with Eq. (15) (based on the fact that the level contours are ellipsoids) is a measure of the circuit robustness. The worst case point $\mathbf{s}_w$ is found by solving:

$$\mathbf{s}_w = \text{argmin}\,(d_w^2 \mid P_{out} = P_{out\,min}). \tag{16}$$

As time t increases, the probabilistic properties of the statistical parameters vector, $\mathbf{s}(t)$, evolve, i.e., the mean $\mathbf{s}_0$ and the covariance matrix $\mathbf{C}$ are functions of time and the worst case distance becomes smaller. In spite of the fact that $\mathbf{s}_0$, the mean, is inside the admissible region, characterized by $P_{out} > P_{out\,min}$, the reliability with

Please cite this article in press as: N. Cucu Laurenciu, S.D. Cotofana, Critical transistors nexus based circuit-level aging assessment and prediction, J. Parallel Distrib. Comput. (2013), http://dx.doi.org/10.1016/j.jpdc.2013.08.006

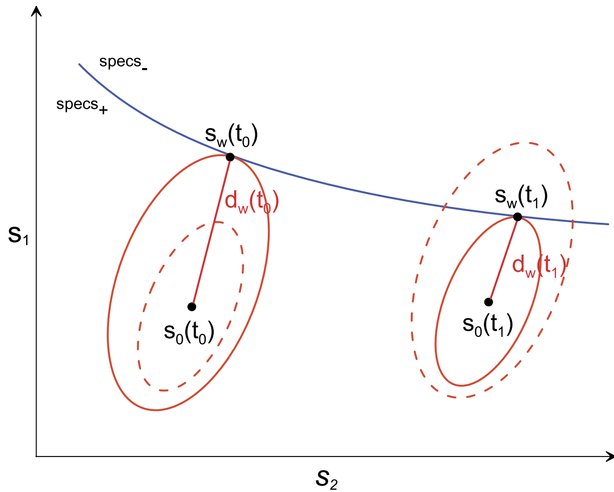
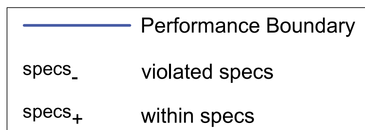


Fig. 2. Graphical representations of lifetime evolution for fixed performance boundary.



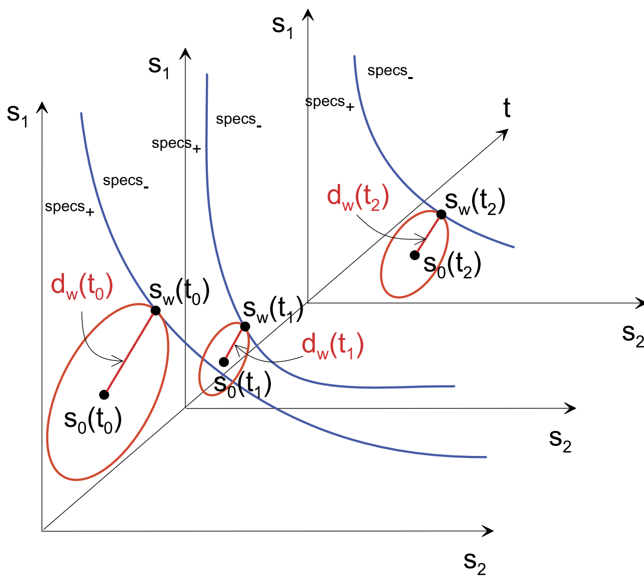


Fig. 3. Graphical representations of lifetime evolution for time varying performance boundary.

the new worst case distance attains its minimum acceptable value and the circuit reaches its end of life. This evolution is graphically captured in Fig. 2.

Actually, the performance boundary defined as $P_{out}(t) = P_{out\,min}$ in the space of $\mathbf{s}$ coordinates, and being characterized by the performance function $P_{out}(\mathbf{d}, \mathbf{s}(t), \mathbf{r})$ which depends on t only via $\mathbf{s}(t)$, could be a too restrictive model for what one might encounter in real situations. For instance, if the range parameters vary in time too, this variation could have remnant – or only transient if the circuit was not exposed for a long time – influence on the physical modifications of the devices. This situation is easier described by allowing the performance boundary to vary in time, as graphically illustrated in Fig. 3. Therefore, we propose to employ a space with one more coordinate – the time – and represent the evolution of the reliability ellipsoids as a tube and the evolution of the performance boundary as a surface exterior to the tube. In this way, both situations are encompassed, that is when the device degradations with T and V_{DD} variations for instance, are remnant, and when they are not remnant.

At this stage we have a model in which the performance scalar P_{out} depends on the vector of statistical parameters $\mathbf{s}$, which has normal and correlated components. Therefore, an orthogonal transform (e.g., KL [7]) can be applied to decorrelate them; after this step the next simplification is to maintain only the two most important components and neglect all the others. It should be noted that these two most important components, retained, are now uncorrelated and as a consequence independent. The following step in developing a workable model is to accept a Markovian evolution and obtain the new values of the two components by applying the transition matrix to the old ones. In fact we deal with two uncorrelated Markov chains, with each component evolving separately. Both processes have a continuous space of states, $\mathbb{R}$, the set of real numbers. In consequence, the probability of a value is obtained by the Chapman–Kolmogorov equation as an integral over $\mathbb{R}$ from the conditional probabilities of that value, given each of the possible previous values:

$$p_{k+1}(y) = \int_{x\in\mathbb{R}} p_k(x) \cdot p_k(y\,|x)\,\mathrm{d}x. \quad (17)$$

In the simplest case, i.e., a stationary Markov process, the model assumes a transition probability that is time independent, that is to say $p_k(y\,|x) = p(y|\,x)$. As previously stated, the evolution of s_1 is independent of s_2. In computing the evolution of the probability density function (pdf) of $s_1(t)$ and $s_2(t)$, we shall replace the continuous time t with a discrete set of integers k. The pdf of $s_1(k+1)$ can be obtained from the pdf of $s_1(k)$ (the same reasoning holds true for s_2) by an integral formula where the Markovian character has to be defined so as to fit the simulation results. This approach is more general than the one developed in [21] and thus can be better fitted to various real conditions.

The two independent Gaussian processes, s_1 and s_2, to which we impose a Markovian character, are therefore Wiener processes. The time evolution of their pdf-s for continuous time is described by [20]:

$$p(s_{i,0}, s_i; t)\mathrm{d}s_i = Prob\left\{s_i < s_i(t) \le s_i + \mathrm{d}s_i \,\middle|\, s_i(0) = s_{i,0}\right\}$$
$$= \frac{1}{\sigma_i\sqrt{2\pi t}} \cdot \exp\left\{-\frac{(s_i - s_{i,0} - \mu_i t)^2}{2\sigma_i^2 t}\right\} \mathrm{d}s_i,$$

where $i \in \{1, 2\}$ and μ_i and σ_i denote the mean and the variance, respectively, of the two processes.

The boundary of the permissible domain in the (s_1, s_2) plane is known and given by the functional relation between P_{out} and the two statistical parameters s_1 and s_2 (see Section 2). As s_1 and s_2 are independent processes, their bi-dimensional pdf is the product of their one dimensional pdf-s. Along the time, the mean (the drift) and variance evolve as for the Wiener process and specifically, increase proportional with t. The circuit starts its life with given values $s_{1,0}$ and $s_{2,0}$ in the admissible domain; as t increases, the mean as well as the variance increase and the point (s_1, s_2) eventually reaches the border. Actually we cannot wait until this event happens: we have to establish the moment when the probability $Prob\,\{P_{out}(s_1, s_2) < P_{min}\}$ and this probability is given by the probability that (s_1, s_2) is out of the border.

In Fig. 4 is presented a sketch of this situation in two successive moments with the 2-dim pdf of (s_1, s_2), while in Fig. 5 is depicted the 2-dim pdf cut away by the performance specifications. Given the Gaussian character of our variables one can compute the probability of the domain D_{ext} (out of the performance border)

Please cite this article in press as: N. Cucu Laurenciu, S.D. Cotofana, Critical transistors nexus based circuit-level aging assessment and prediction, J. Parallel Distrib. Comput. (2013), http://dx.doi.org/10.1016/j.jpdc.2013.08.006

ARTICLE IN PRESS

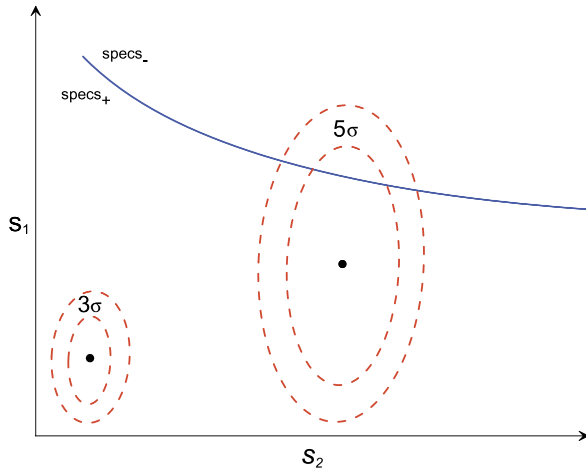


Fig. 4. 2D-PDF evolution.

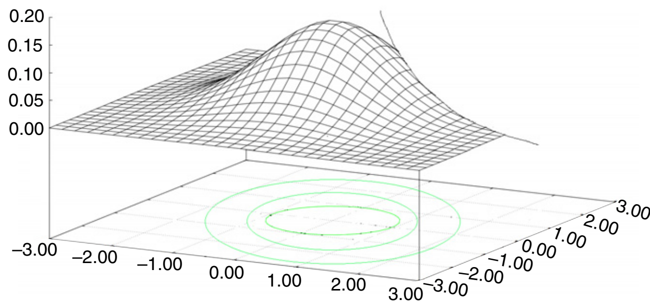


Fig. 5. 2D-PDF truncated by performance boundary.

at any moment in time. For an estimation of the moment we are interested in, it is enough to have the values of:

$$Q(t) = Prob\,\{(s_1(t),\ s_2(t)) \in D_{ext}\} \tag{18}$$

in a finite number of moments and use a linear interpolation between them. In this way (the intersection of $Q(t)$ with the horizontal line $Q = Q_{max}$ admissible) we obtain the moment when the circuit reaches its end-of-life.

We stress out that using the hypotheses mentioned above we are able to bypass the difficulties of a direct Markovian model (in our model the Markovian character is included in the Wiener model for which there are classical results). The parameters of the Wiener processes have to be obtained by means of simulations.

When the **r** parameters are varying too, it is necessary to move the border accordingly at the same time as the pdf of $(s_1,\ s_2)$ is evolving in the $(s_1,\ s_2)$ space. There are two situations: either one knows their variation or only a pdf of this border (and so of the domain D_{ext}), is known. In the last case we have to compute $Q(t)$ for any position of the border – we shall index the possible positions at time t by a variable u – and to obtain the probability we look for as a weighted value of the probabilities for each D_{ext}:

$$Q(t) = \int Q(t,\ u) \cdot Prob\,\{D_{ext}(u)\}\,\mathrm{d}u. \tag{19}$$

It is very likely that we have only a few values of the **r** parameters, as for instance three values of V_{DD} with probabilities p_1, p_2, and p_3; for a time t. In such a case $Q(t)$ can be obtained as:

$$Q(t) = p_1 \cdot Q(t,\ D_{ext1}) + p_2 \cdot Q(t,\ D_{ext2}) + p_3 \cdot Q(t,\ D_{ext3}), \tag{20}$$

where $p_1 + p_2 + p_3 = 1$. We note inhere that this formulation does not contain the case when a variation of the **r** parameters induces modifications on other parameters of the function P_{out}. In such a case, the Wiener process model parameters have to be continuously adapted at run-time.

5. Performance evaluation

In this section, the framework with fixed topology-dependent weights and the Markovian circuit-level aging framework are validated and their end-of-life estimation accuracies are evaluated. The simulation is conducted on the ISCAS-85 c499 circuit, which is a single-error-correcting circuit with 41 inputs, 32 outputs, and 202 gates, using PTM 45 nm technology [5]. The reliability analysis (BTI and HCI aging) is carried in Cadence RelXpert and Virtuoso Spectre simulators [6], using the AgeMOS model extracted in BSIMPro+ for PTM 45 nm technology [22]. The transistor-level static timing analysis in performed in Synopsys Pathmill [28].

The validity of estimating a circuit end-of-life from the end-of-life of the critical transistors in the kernel set, is examined by exposing the circuit to several stress profiles (e.g., varying duty-cycle, temperature, input vectors). Based on each profile's fresh and aged timing reports, we determine the set of aging critical paths, i.e., we select the paths with propagation delay exceeding the clock period. In our case we impose an end-of-life target of 10% propagation delay degradation, and retain the first 100 critical paths. The initial set of transistors that constitute the 100 critical paths and which is to be reduced to a set of critical ones, consists of 53 transistors. Then, according to the methodology described in Section 3, the regression matrix is derived, and implicitly the reduced set of critical transistors. Fig. 6 illustrates the regression matrix obtained for the analyzed circuit. The input and output variables are the end-of-life of the critical transistors and the end-of-life of the critical paths, respectively, obtained from simulation. Based on the input and output variables, the regression coefficients, i.e., the topology dependent weights, are obtained using the model from Section 3. In the left subfigure, the input variables that are discarded from the model are represented in black, while the reduced set of inputs – in our case 15 from a total of 53 – that are relevant for synthesizing the output responses – in our case 100 aging critical paths – are represented in white. The right subfigure depicts in grayscale the variable regression coefficients w_{ij} corresponding to each relevant input variable, for all the output responses.

Having determined the minimum-size kernel of critical transistors and their topology dependent coefficients, we are now in the position to validate the resulted model for a new set of input aggression profiles, using Eq. (10). Fig. 7 illustrates the normalized simulated circuit end-of-life values vs. the normalized estimated circuit end-of-life values in the case of the new set of input aggression profiles. The simulation results reveal a mean estimation error of 15% and a variance of 6%, which confirms that the determined kernel of critical transistors can be utilized to estimate the circuit end-of-life at run-time fairly accurate. A remark is in order: To achieve a good estimation of a circuit end-of-life, besides the matter of choice of solving the regression problem, the initial sampling for multiple levels of stress should be carefully considered.

Since the reliability aware management of integrated circuits implemented in advanced technology nodes requires reasonably accurate but fast run-time reliability profiling, a further reduction of the number of aging measurement sites could be desired. To this extent, we study the trade-offs between the number of critical transistors that are used for end-of-life circuit estimation, and the circuit end-of-life estimation accuracy. Fig. 8 depicts the error analysis of the circuit end-of-life, for different subsets – with different cardinality – of critical transistors, when subjecting the circuit to 5 new stress profiles. For each stress profile, 5 subsets

Please cite this article in press as: N. Cucu Laurenciu, S.D. Cotofana, Critical transistors nexus based circuit-level aging assessment and prediction, J. Parallel Distrib. Comput. (2013), http://dx.doi.org/10.1016/j.jpdc.2013.08.006

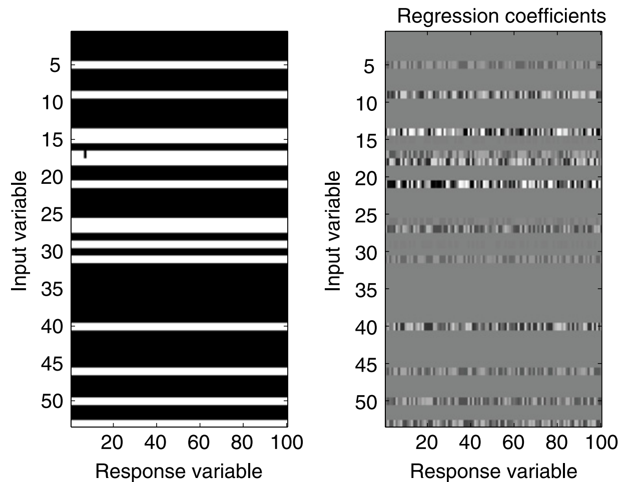


Fig. 6. The regression coefficients determining the reduced set of critical transistors.

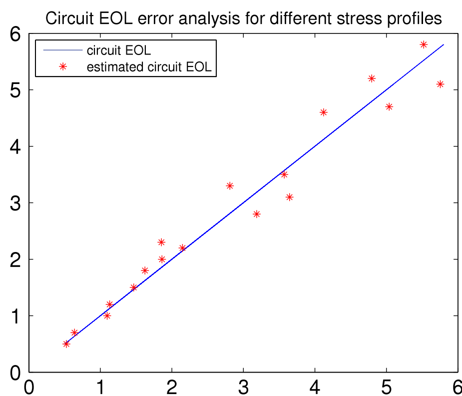


Fig. 7. Error analysis of circuit end-of-life estimation based on the end-of-life values of the critical transistors.

of critical transistors with different cardinalities, which are obtained by reducing the initial critical transistors kernel with 2%, 5%, 10%, 25%, and 30%, are being considered. The percentage of estimation accuracy loss is reported relative to the estimation accuracy obtained when using the entire kernel of critical transistors. The transistors are eliminated based on their relevance in estimating the circuit end-of-life (i.e., the less relevant goes out first). We observe a similar trend of the end-of-life circuit estimation quality loss when decreasing the number of critical transistors for all considered stress profiles. As concerns the differences in the rate of estimation accuracy loss, they can be attributed to the relevance of the dropped transistors in estimating the model responses for considered input stress profiles. However, taking into consideration that in most situations a very precise estimation of the circuit end-of-life is not required, a coarse reliability assessment is sufficient to enable graceful performance degradation and prolong the circuit lifetime via aging mitigation and compensation techniques. One can observe in Fig. 8 that for the considered circuit, up to 30% area overhead reduction (5 sensors instead of 15 to monitor the reliability of a 202 gates circuit) can be achieved for less than 18% loss in circuit end-of-life estimation accuracy (reported relative to the estimation accuracy achieved by employing the entire kernel of critical transistors), which makes it a potentially feasible approach for practical implementations.

For a more accurate estimation of the circuit end-of-life, that takes into account the history of aging, we apply the Markovian

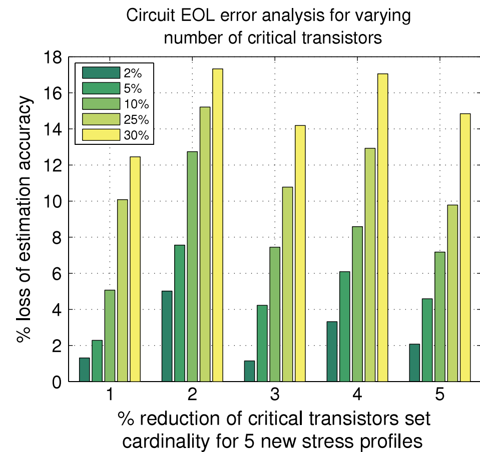


Fig. 8. Error analysis of circuit end-of-life estimation based on the end-of-life values of the critical transistors.

framework on the kernel of critical transistors previously validated. For the purpose of illustration, we employ Monte Carlo simulation loops, approach which is typical for analog circuits, where the analytical expressions of circuit performance features as functions of statistical parameters are not known. We choose as circuit performance metric the propagation delay. As concerns the statistical parameters, we use the threshold voltage, Vth, the low-field mobility μ_0, the oxide thickness t_{ox}, and the oxide capacitance C_{ox}. After decorrelation, the components Vth and μ_0 are retained.

In Fig. 9, is depicted the normalized circuit end-of-life, which is defined as the time when the propagation delay is degraded by υ%. For expository purposes we define the end-of-life target for the considered simulation framework as $\upsilon = 10\%$ degraded propagation delay. We consider several stress profiles (e.g., varying duty-cycle, temperature, input vectors), and obtain the corresponding performance boundary for defined end-of-life target in the $(\mathbf{V}_{th}, \boldsymbol{\mu}_\mathbf{0})$ space, as result of reliability analysis (NBTI and HCI aging) and Monte Carlo simulation. For each profile and corresponding data set of statistical parameters, we determine $Q(t)$ in a finite number of moments, interpolate them and estimate the end-of-life time moment. This is compared against the accurate end-of-life value which is obtained by means of simulation, i.e., the time moment when $\mathbf{s} = \mathbf{s_w}$, for the obtained performance boundary. Fig. 9 illustrates the obtained circuit end-of-life prediction accuracy using the Markovian framework on the set of critical transistors. We obtained an approximation error with mean ($<$10%) and standard deviation ($<$15%). As expected, the estimated end-of-life values are further from the values obtained with Cadence. We attribute this to the Markovian approach and to the fact that we use multiple monitors to quantify the aging process. In fact, as the Markovian model takes into consideration more parameters and aging sources, these estimated end-of-life values may be closer to the real end-of-life values but for the time being we do not have the means to validate this conjecture. However, the proposed Markovian framework necessitates the monitoring of multiple degradation parameters per transistor, e.g., Vth, μ, and hence multiple sensors are required for one transistor. This makes this approach less feasible for run-time aging assessment and prediction, and better suited at design-time, enabling a robust, fast and accurate aging evaluation, which takes into account the history of the degradation caused by joint effects of process, environmental, and aging-induced variations.

Furthermore, the proposed Markovian framework is general, and hence suitable for emergent nanoscale technologies, under the provision that the technology is known (and implicitly the afferent design and statistical parameters are known). In this work,

Please cite this article in press as: N. Cucu Laurenciu, S.D. Cotofana, Critical transistors nexus based circuit-level aging assessment and prediction, J. Parallel Distrib. Comput. (2013), http://dx.doi.org/10.1016/j.jpdc.2013.08.006

ARTICLE IN PRESS

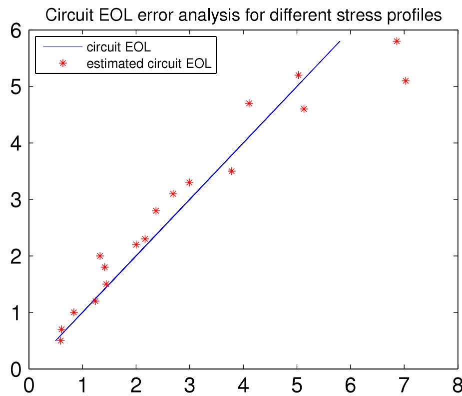


Fig. 9. Error analysis of circuit end-of-life estimation, using the Markovian statistical framework on the set of critical aging transistors.

we particularized the framework for the bulk CMOS in 45 nm technology node, using certain commonly employed parameters such as Vth as statistical parameter. However the same line of reasoning can be applied for newer technologies to similar or different electrical and/or performance parameters. As far as the topology dependent framework is concerned, it can also be applied for emergent nanotechnologies, even if the circuit critical paths may have different constituent blocks instead of the bulk CMOS transistors that we employed for expository purposes.

6. Conclusions

In this paper we proposed to estimate a circuit end-of-life based on a minimum-size set of critical transistors. After the selection at design time of the optimal critical transistors and their topology dependent weights, a circuit end-of-life can be estimated at run time as the sum of the measured end-of-life values of the critical transistors weighted by the design-time determined topology dependent coefficients. Based on the same set of critical transistors, an alternative way to determine the circuit end-of-life with increased accuracy is presented via the proposed Markovian aging framework which adapts in time its performance boundary. We noticed that, if the circuit has been stressed only for a short duration in time, the damage on the physical and electrical parameters might not be permanent. This is opposed to the case when the circuit is exposed to stress conditions for a long period of time and as a consequence exhibits remnant parameters drifts. We proposed to account for these two cases by adapting the performance boundary according to the variation of statistical parameters. In this way we obtained an aging framework that is more flexible and thus better fitted to real conditions. Furthermore, since the age at time t is a functional, depending on the entire set of values taken by the degradation parameters starting from time 0 till time t, we imposed a Markovian character to our circuit level age prediction model by taking into account the history of the statistical parameters that fluctuate due to process variations and dynamic operating conditions. We validated the proposed framework for different aggression profiles, for the ISCAS-85 c499 circuit, implemented in PTM 45 nm. We obtained a kernel of 15 critical transistors from a total of 1526 transistors. Experimental results for the 15 critical transistors kernel indicate a small approximation error (i.e., mean smaller than 15% and standard deviation smaller than 6%) for the c499 circuit end-of-life (EOL), quantitatively confirming the accuracy of the lifetime evaluation. Furthermore, as the area overhead is determined by the number of critical transistors, we also investigated the implications of reducing their number on the reliability assessment accuracy. When only 5 transistors are included into the critical set instead of 15, which results in a 66% area overhead reduction, the EOL estimation accuracy diminished with 18%. This indicates that area vs. accuracy trade-offs are possible, while maintaining the aging prediction accuracy within reasonable bounds.

References

[1] B.E. Barrett, J.B. Gray, A computational framework for variable selection in multivariate regression, Stat. Comput. 4 (1994) 203–212.
[2] S. Bhardwaj, W. Wang, R. Vattikonda, Y. Cao, S. Vrudhula, Predictive modeling of the NBTI effect for reliable design, in: Design, Automation and Test in Europe, 2006, pp. 189–192.
[3] S. Borkar, Designing reliable systems from unreliable components: the challenges of transistor variability and degradation, in: Micro, 2005, pp. 10–16.
[4] L. Breiman, J.H. Friedman, Predicting multivariate responses in multiple linear regression, J. R. Stat. Soc. Ser. B (Methodological) (1997) 3–54.
[5] http://www-device.eecs.berkeley.edu/bsim/.
[6] http://www.cadence.com/us/pages/default.aspx.
[7] K.R. Castleman, Digital Image Processing, Pretice Hall, 1996.
[8] R. Cranwell, Ground vehicle reliability design-for-reliability, in: DoD Maintenance Symposium, Orlando, FL, 2007.
[9] N. Cucu Laurenciu, S.D. Cotofana, A Markovian, variation-aware circuit-level aging model, in: IEEE/ACM International Symposium on Nanoscale Architectures, 2012.
[10] N. Cucu Laurenciu, S.D. Cotofana, Context aware slope based transistor-level aging model, Microelectron. Reliab. 52 (2012) 9–10.
[11] M. Dietrich, J. Haase, Process Variations and Probabilistic Integrated Circuit Design, Springer, 2012.
[12] K. Eshbaugh, Generation of correlated parameters for statistical circuit simulation, IEEE Trans. Comput.-Aided Des. (1992) 1198–1206.
[13] R. Huang, H.M. Wu, J.F. Kang, et al., Challenges of 22 nm and beyond CMOS technology, Sci. China (2009) 1491–1533.
[14] Process integration, devices, and structures, in: International Technology Roadmap for Semiconductors, 2009.
[15] S. Kumar, C.H. Kim, S. Sapatnekar, An analytical model for negative bias temperature instability, in: International Conference on Computer-Aided Design, 2006.
[16] S. Kumar, J. Li, C. Talarico, J. Wang, A probabilistic collocation method based statistical gate delay model considering process variations and multiple input switching, in: Design, Automation and Test in Europe, 2005, pp. 770–775.
[17] X. Li, Asymptotic probability extraction for non-normal distribution of circuit, in: IEEE/ACM International Conference on Computer Aided Design, 2004, pp. 2–9.
[18] H. Liu, M. Palatucci, J. Zhang, Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery, in: International Conference on Machine Learning, 2009.
[19] Y. Lu, L. Shang, H. Zhou, H. Zhu, F. Yang, X. Zeng, Statistical reliability analysis under process variation and aging effects, in: 46th ACM/IEEE Design Automation Conference, DAC'09, 2009, pp. 514–519.
[20] J. Medhi, Stochastic Processes, New Age International Limited, Publishers, 2002.
[21] X. Pan, H. Graeb, Lifetime yield optimization of analog circuits considering process variations and parameter degradations, in: Advances in Analog Circuits, InTech, 2011, pp. 56–99.
[22] http://ptm.asu.edu/.
[23] S. Sapatnekar, Timing, Kluwer Academic Publishers, 2004.
[24] T. Siddiqua, S. Gurumurthi, M.R. Stan, Modeling and analyzing NBTI in the presence of Process Variation, in: International Symposium on Quality Electronic Design, 2011, pp. 514–519.
[25] T. Simila, J. Tikka, Input selection and shrinkage in multiresponse linear regression, Comput. Statist. Data Anal. (2007) 406–422.
[26] Alvin W. Strong, et al., Reliability Wearout Mechanisms in Advanced CMOS Technologies, John Wiley and Sons, New Jersey, 2009.
[27] A.W. Strong, E.Y. Wu, R.P. Vollertsen, S. Sune, et al., Reliability Wearout Mechanisms in Advanced CMOS Technologies, John Wiley and Sons, Inc, Hoboken, New Jersey, 2009.
[28] http://www.synopsys.com/home.aspx.
[29] J.A. Tropp, Algorithms for simultaneous sparse approximation, part ii: Convex relaxation, Signal Process. 86 (2006) 589–602.
[30] B.A. Turlach, W.N. Venables, S.J. Wright, Simultaneous variable selection, Technometrics (2005) 349–363.
[31] B. Vaidyanathan, A.S. Oates, Y. Xie, Y. Wang, NBTI-aware statistical circuit delay assessment, in: International Symposium on Quality Electronic Design, 2009, pp. 13–18.
[32] Y. Wang, S. Cotofana, L. Fang, A novel virtual age reliability model for time-to-failure prediction, in: IEEE International Integrated Reliability Workshop Final Report, 2010, pp. 102–105.
[33] Y. Wang, S. Cotofana, L. Fang, A unified aging model of nbti and hci degradation towards lifetime reliability management for nanoscale MOSFET circuits, in: IEEE/ACM International Symposium on Nanoscale Architectures, 2011, pp. 175–180.
[34] W. Wang, V. Reddy, B. Yang, V. Balakrishnan, S. Krishnan, Y. Cao, Statistical prediction of circuit aging under process variations, in: 12th International Symposium on Quality Electronic Design, ISQED, 2008, pp. 13–16.

Please cite this article in press as: N. Cucu Laurenciu, S.D. Cotofana, Critical transistors nexus based circuit-level aging assessment and prediction, J. Parallel Distrib. Comput. (2013), http://dx.doi.org/10.1016/j.jpdc.2013.08.006

[35] M. Yuan, Y. Lin, Model Selection and Estimation in Regression with Grouped Variables. Technical Report, University of Wisconsin, 2004.

Nicoleta Cucu Laurenciu received her M.Sc. degree in Computer Engineering from Delft University of Technology, The Netherlands, in 2010, and she is currently a Ph.D. Candidate in the Computer Engineering Laboratory, Delft University of Technology, the Netherlands. Her research interests are in the area of reliability of nanoelectronic devices and systems, and Dynamic Reliability Management of multi/many-core architectures.

Sorin D. Cotofana received the M.Sc. degree in Computer Science from the "Politehnica" University of Bucharest, Romania, and the Ph.D. degree in Electrical Engineering from Delft University of Technology, The Netherlands. He is currently an Associate Professor with the Electrical Engineering, Mathematics and Computer Science Faculty, Delft University of Technology, Delft, the Netherlands. His current research is focused on: (i) the design and implementation of dependable/reliable systems out of unpredictable/unreliable components; (ii) aging assessment/prediction and lifetime reliability aware resource management; and (iii) unconventional computation paradigms and computation with emerging nano-devices. He is a HiPEAC member, a senior IEEE member (Circuits and System Society (CASS) and Computer Society), Chair of the Giga-Nano IEEE CASS Technical Committee, and IEEE Nano Council CASS representative.

Please cite this article in press as: N. Cucu Laurenciu, S.D. Cotofana, Critical transistors nexus based circuit-level aging assessment and prediction, J. Parallel Distrib. Comput. (2013), http://dx.doi.org/10.1016/j.jpdc.2013.08.006