Statistical Reliability Analysis Under Process Variation and Aging Effects

Yinghai Lu[1], Li Shang[2], Hai Zhou[1,3], Hengliang Zhu[1], Fan Yang[1], Xuan Zeng[1*]
[1]State Key Lab of ASIC & System, Microelectronics Dept., Fudan University, China
[2]ECEE, University of Colorado, Boulder, U.S.A., [3]EECS, Northwestern University, U.S.A.

***Abstract*—Circuit reliability is affected by various fabrication-time and run-time effects. Fabrication-induced process variation has significant impact on circuit performance and reliability. Various aging effects, such as negative bias temperature instability, cause continuous performance and reliability degradation during circuit run-time usage. In this work, we present a statistical analysis framework that characterizes the lifetime reliability of nanometer-scale integrated circuits by jointly considering the impact of fabrication-induced process variation and run-time aging effects. More specifically, our work focuses on characterizing circuit threshold voltage lifetime variation and its impact on circuit timing due to process variation and the negative bias temperature instability effect, a primary aging effect in nanometer-scale integrated circuits. The proposed work is capable of characterizing the overall circuit lifetime reliability, as well as efficiently quantifying the vulnerabilities of individual circuit elements. This analysis framework has been carefully validated and integrated into an iterative design flow for circuit lifetime reliability analysis and optimization.**

Categories and Subject Descriptors:
J.6 [Computer-Aided Engineering]: Computer-Aided Design
B.8.1 [Performance and Reliability]: Reliability, Testing, and Fault-Tolerance
General Terms: Design, Algorithms, Performance
Keywords: NBTI, Yield, Process variations

I. Introduction

Aggressive scaling of CMOS process technology poses serious challenges on the lifetime reliability of integrated circuits (ICs). IC lifetime reliability is affected by fabrication-induced process variation and run-time aging effects [3]. Feature size reduction increases the difficulty of precise fabrication process control. Fabrication induced geometric and electrical parameter variations, e.g., changes in device effective channel length and threshold voltage, have significant impact on IC performance and reliability. Meanwhile, run-time aging effects, such as electromigration, thermal cycling, and negative bias temperature instability (NBTI), have become another fast-growing concern of IC lifetime reliability. NBTI is known to be the dominating circuit lifetime aging effect [12], [7]. The occurrence of NBTI is due to the generation of traps at $Si\text{-}SiO_2$ interface when PMOS devices are negatively stressed, e.g., $V_{gs} = -V_{dd}$. This effect causes temporal increase of PMOS threshold voltage (V_{th}) and long-term performance degradation.

Process variation [16], [15] and NBTI effect [13], [6], [14], [21], [26], [20] have both drawn significant attention in the recent past. Most of the past work treats them as two independent issues, and addresses the impact of each effect on IC reliability and performance separately. However, IC reliability is jointly affected by both effects. In addition, process variations and NBTI effect have strong influence on each other. As reported by Bhardway et al., NBTI-induced threshold voltage shift of PMOS transistor depends not only on its working condition (such as input duty cycle and temperature), but also on the underlying process parameters such as original threshold voltage and dioxide thickness [6]. Due to fabrication-induced process variations, NBTI effect shall be modeled as a random process. On the other hand, the circuit timing statistics will be affected by the NBTI effect as well as the process variation as time evolves.

Recently, work starting to consider both effects has been reported. In [11], NBTI effect (ΔV_{th}) is modeled as a random process. This work, however, ignores the variation of other process parameters. In [4], the authors consider both process variation and NBTI effect for standard cell modeling and optimization. Process parameters are treated as random variables and modeled with response surface method. In this work, however, NBTI effect is incorporated with the worst-case delay model. Since NBTI effect is a strong function of circuit run-time condition, the worst-case approximation is pessimistic (up to $30\times$ increase of ΔV_{th} versus the nominal has been reported in [4]). In [22], the authors present a thorough analysis of circuit aging under the variation of threshold voltage. This work's main focus is on single path circuit modeling. A comprehensive analysis of IC performance and reliability yet requires statistical techniques to address correlated paths and other process parametric variations. We believe this work starts an important direction, and our study will expand and build up on this foundamental work.

In this article, we present an analysis framework to evaluate the IC lifetime reliability by jointly considering process variation and NBTI effect. This work makes the following contributions:

- We present a nonlinear scalable statistical gate delay aging model, which considers both run-time gate working condition and fabrication-induced process variation.
- We present a statistical timing analysis framework using the proposed gate delay model, which is capable of characterizing the performance and reliability degradation under process variation and run-time aging. A fast pruning algorithm is proposed to improve the analysis efficiency.
- We present a criticality and sensitivity analysis method to quantify the reliability impact of each individual circuit element. Such quantification enables efficient iterative IC reliability optimization flow.

The rest of the article is organized as follows. Section II introduces the proposed analysis framework. Sections III, IV, and V describe in detail the proposed modeling, analysis, and optimization methods. Section VI reports the experimental results. The paper is concluded in Section VII.

II. Overview of Aging-Aware Statistical Framework

The proposed aging-aware statistical timing analysis framework is shown in Figure 1. The analysis framework consists of three key components:

Gate-Level Aging-Aware Statistical Timing Model: Given a technology library in the SPICE netlist form, it characterizes the process

[*] Corresponding author. E-mail: xzeng@fudan.edu.cn

DAC 2009, July 26 - 31, 2009, San Francisco, California, USA
Copyright 2009 ACM ACM 978-1-60558-497-3 -6/08/0006 ...$5.00.

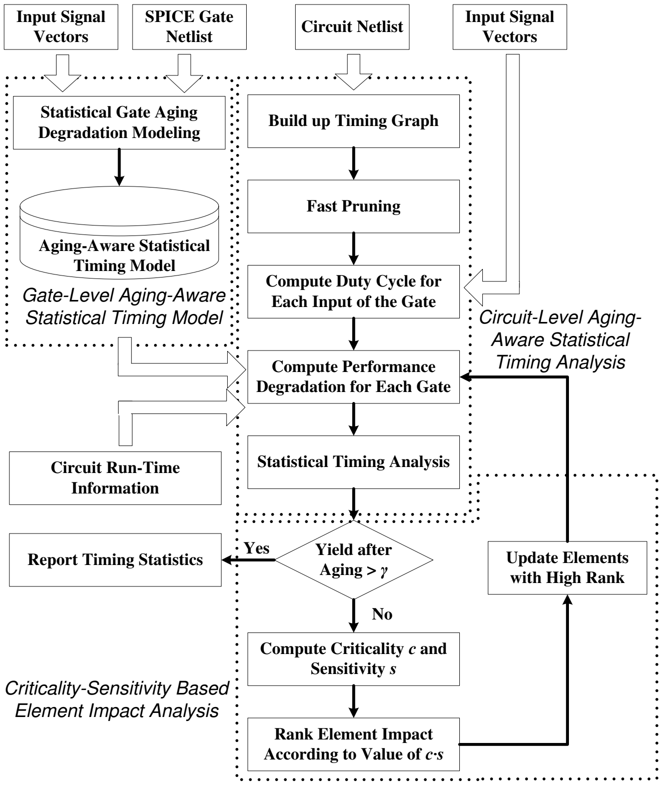
Input Signal Vectors
SPICE Gate Netlist
Circuit Netlist
Input Signal Vectors
Statistical Gate Aging Degradation Modeling
Aging-Aware Statistical Timing Model
Gate-Level Aging-Aware Statistical Timing Model
Build up Timing Graph
Fast Pruning
Compute Duty Cycle for Each Input of the Gate
Circuit-Level Aging-Aware Statistical Timing Analysis
Compute Performance Degradation for Each Gate
Circuit Run-Time Information
Statistical Timing Analysis
Report Timing Statistics
Yes
Yield after Aging > γ
No
Update Elements with High Rank
Compute Criticality c and Sensitivity s
Criticality-Sensitivity Based Element Impact Analysis
Rank Element Impact According to Value of c·s

Fig. 1. Aging-aware statistical framework.

variation and NBTI aging effect of each gate element, and generates a statistical model in polynomial chaos expansion (PCE) form incorporating NBTI-induced aging degradation and process parameter variation. The NBTI aging model considers various run-time factors, such as temperature, input signal probability, and duty cycle.

Circuit-Level Aging-Aware Statistical Timing Analysis Method: It conducts statistical timing analysis on the entire circuit using the proposed gate-level model. Logic simulation is first conducted to collect the signal duty cycle information of each logic element. Given process variations and run-time circuit conditions, such as voltage and temperature, gate performance degradation is expressed in a unified PCE form, which is used by statistical static timing analysis (SSTA) to compute the performance degradation of the entire circuit. A circuit pruning algorithm is developed, which incrementally removes gates and edges with sufficiently large time slacks during analysis, thereby improving analysis efficiency without sacrificing accuracy.

Criticality-Sensitivity Based Element Vulnerability Analysis Method: It conducts circuit criticality and sensitivity analysis in order to quantify the impact of individual elements on circuit lifetime timing improvements. This analysis provides guidance to IC aging-aware reliability optimization, and enables an efficient iterative IC reliability analysis and optimization framework.

The following sections discuss the details of the proposed modeling and analysis methods.

III. Statistical Gate Aging Modeling

A. Parametric NBTI Aging Modeling

This section describes a parametric method to model PMOS NBTI effect, which extracts and formulates NBTI run-time dependencies, e.g., temperature and signal probability, in a compact form that allows rapid estimation of NBTI-induced time degradation under arbitrary run-time conditions. This proposed method can also be extended to model other lifetime aging effects. The NBTI effect manifests itself as increase of PMOS threshold voltage and degradation of circuit timing. NBTI physical mechanism has been studied in [7], [19]. A NBTI model under arbitrary dynamic temperature variation is proposed in [26]. In [6], the authors propose a long-term NBTI model, which provides an analytical upper bound estimation of the NBTI impact over time, described as follows.

$$\Delta V_{th}(t) = \left(\frac{\sqrt{K_v^2 \alpha T_{clk}}}{1 - \beta_t^{1/2n}} \right)^{2n} \tag{1}$$

where

$$K_v = \left(\frac{q t_{ox}}{\epsilon_{ox}} \right)^3 K^2 C_{ox} (V_{gs} - V_{th}) \sqrt{C} \exp\left(\frac{2E_{ox}}{E_o} \right) \tag{2}$$

$$C = T_o^{-1} \exp\left(-\frac{E_\alpha}{kT} \right) \tag{3}$$

$$\beta_t = 1 - \frac{2\xi_1 t_e + \sqrt{\xi_2 C (1-\alpha) T_{clk}}}{2t_{ox} - \sqrt{Ct}} \tag{4}$$

where α is the average signal duty cycle, T is the average temperature, V_{th} is the initial threshold voltage and t_{ox} is thickness of gate dielectric. For the sake of simplicity, please refer to [6] for detailed explanation of other parameters.

This model shows that the NBTI effect is a strong function of the run-time temperature T and signal probability (duty-cycle) α of the logic gate. In this work, we extract the dependency on T and α from the long-term model described in Equation (1). We further assume T_{clk} is small based on the fact that modern high-speed IC designs are typically clocked at the multi-gigahertz range. Using the similar reduction technique as in [20] and considering the exponential temperature dependency shown in Equation (3), the NBTI-induced threshold shift model can be approximately reduced to

$$\Delta V_{th}(T, \alpha, t) = b e^{-\frac{nE_\alpha}{kT}} \left(\frac{\alpha}{1-\alpha} \right)^n t^n \tag{5}$$

where k is Boltzmann constant, $E_\alpha = 0.49eV$ and b is a fitting constant. Figure 2 demonstrates the relative error of our simplified model in Equation 5 under varying working conditions against the original long-term model in Equation 1, using the 65nm CMOS technology. The temperature ranges from $320K$ to $380K$ and the average duty cycle ranges from 0.1 to 0.95. As is displayed in the figure, the simplified model achieves very good accuracy within normal work conditions.

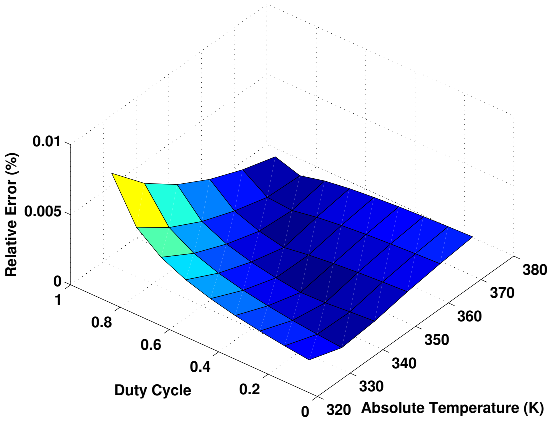

Fig. 2. Relative Error of Model in Equation 5 Against the Long-term Model (Equation 1).

Following the alpha-power law [18], the first-order gate delay can be approximated as a linear function of the threshold voltage. The gate delay can then be expressed as follows [20], [23]:

$$\Delta d(T, \alpha, t) = \tilde{b} e^{-\frac{nE_\alpha}{kT}} \left(\frac{\alpha}{1-\alpha} \right)^n t^n \tag{6}$$

where constant $\tilde{b}$ can be fitted from SPICE characterization. A primary advantage of using Equation (6) to characterize gate-level NBTI effect is that, given a reference model pre-characterized at T_{ref} and α_{ref}, the aging effect of a logic gate under any arbitrary T and α can be efficiently calculated using parameter scaling, which will be discussed in detail in Section IV-A.

B. Considering Process Variations in NBTI Aging Modeling

The NBTI-induced gate delay degradation depends not only on circuit run-time conditions but also on fabrication-determined process parameters, such as the initial threshold voltage V_{th} and oxide thickness t_{ox}. Due to fabrication-induced process variations, the NBTI aging process and its impact on circuit timing become a random process. To model NBTI-induced aging under process variation, we apply the stochastic collocation method, which was originally proposed to model the gate delay under uncertainty [15]. Considering a set of process parameters as a random variable vector with normal distribution

$$\vec{\xi} = [L, V_{th}, t_{ox}, \ldots], \tag{7}$$

the variation of gate delay degradation can be modeled using the polynomial chaos expansion on the set of random variables.

$$\Delta d(\vec{\xi}, T, \alpha, t) \approx \sum_{j}^{P} c_j \Phi_j(\vec{\xi}) e^{-\frac{nE_\alpha}{kT}} \left(\frac{\alpha}{1-\alpha}\right)^n t^n \tag{8}$$

where $\{\Phi_j\}_{j=1}^{P}$ is the complete set of the d-dimension Hermite polynomials up to the l-th order, and c_j is the unknown coefficient. Hermite polynomials form a set of orthogonal basis of Hilbert space under Gaussian measure, and thus is the best if $\vec{\xi}$ is Gaussian. Second order Hermite polynomials are sufficient in practice. The coefficients c_j can be determined using the stochastic collocation method [5], [15]. Using Equation (8), the delay of a logic gate at time t can then be expressed as follows.

$$d(\vec{\xi}) = d_0(\vec{\xi}) + \Delta d(\vec{\xi}, T, \alpha, t) \tag{9}$$

where $d_0(\vec{\xi})$ is the initial gate delay after chip fabrication. $\Delta d(\vec{\xi}, T, \alpha, t)$ represents the time-dependent gate aging effect. $d_0(\vec{\xi}) = \sum_j^P d_j \Phi_j(\vec{\xi})$ is expressed in polynomial chaos form with respect to the same set of random variables as Δd, and the coefficients are computed using the same stochastic collocation method. For gates with multiple inputs, this model is applied for each input of the gate.

IV. Aging-Aware Statistical Timing Analysis

Given a circuit netlist and the aging-aware variational gate delay model (Equation 9) as inputs, the proposed aging-aware statistical timing analysis method computes the aging effect of each logic gate based on its run-time condition, and carries out circuit-level statistical timing analysis.

A. Computation of Gate NBTI Aging Effect

For each type of logical gate provided by the technology library, the NBTI aging effect is characterized once under a reference working condition T_{ref} and α_{ref}, and is expressed as $\Delta d(\vec{\xi}, T_{ref}, \alpha_{ref}, t)$. During circuit-level logic analysis, for each logic gate i, the duty cycle α_i of its input signal is estimated by circuit simulation using user-provided input signal vectors, and the gate temperature T_i is provided from chip thermal profile. The aging effect of gate i can then be calculated by scaling the referenced model as follows.

$$\Delta d(\vec{\xi}, T_i, \alpha_i, t) = \Delta d(\vec{\xi}, T_{ref}, \alpha_{ref}, t) \cdot R_T \cdot R_\alpha \tag{10}$$

where the scaling factors for temperature T_i and signal duty cycle α_i are

$$R_T(T_{ref}, T_i) = exp(\frac{nE_\alpha}{k} \cdot \frac{T_i - T_{ref}}{T_{ref}T_i}) \tag{11}$$

and

$$R_\alpha(\alpha_{ref}, \alpha_i) = \left[\frac{\alpha_i(1-\alpha_{ref})}{\alpha_{ref}(1-\alpha_i)}\right]^n \tag{12}$$

Due to the exponential dependence of the temperature and error introduced during model simplification, a maximum of $\pm 25K$ temperature difference is allowed in order to achieve good accuracy of the scaled aging effect for a gate (This setting is used in the experimental result section.). In order to cover the complete circuit operation range, we develop piece-wise gate aging model. In addition, the above scaling model is not applicable when $\alpha = 1$, which indicates that the gate is under static stress. Using the static NBTI model, a scaling aging model can also be developed for $\alpha = 1$ using the same method described above.

B. Equivalent Aging Time Analysis

Circuit workload may vary over time, so is the aging process. For example, as shown in Figure 3, a gate experiences three different working conditions, (T_1, α_1), (T_2, α_2) and (T_3, α_3), with a duration of t_1, t_2 and t_3, respectively. In this work, we introduce *equivalent aging time* to facilitate characterization of the aging effect under such varying conditions. For the sake of clarity, the dependency of the aging effect on $\vec{\xi}$ is omitted. Given the aging effect of a logic gate under condition (T_2, α_2), the equivalent aging effect under condition (T_1, α_1) is described as follows.

$$\Delta d(T_2, \alpha_2, t_{eqv1}) = \Delta d(T_1, \alpha_1, t_1) \tag{13}$$

Using Equation 6, the equivalent aging time t_{eqv1} can be computed as

$$t_{eqv1} = t_1 \cdot [R_T(T_2, T_1) R_\alpha(\alpha_2, \alpha_1)]^{1/n} \tag{14}$$

where R_T and R_α are defined in Equation (11) and (12). Then, the aging effect of the gate at t_2 equals that of the gate working under (T_2, α_2) during time $(0, t_{eqv1} + t_2)$, and can be computed as $\Delta d(T_2, \alpha_2, t_{eqv1} + t_2)$. This procedure can be carried out inductively whenever the working condition changes.

Figure 3 demonstrates the use of equivalent aging time to estimate the overall aging process under three different working conditions. The dotted lines are the equivalent aging times computed at each transition using Equation (14). The solid lines are actual aging durations.

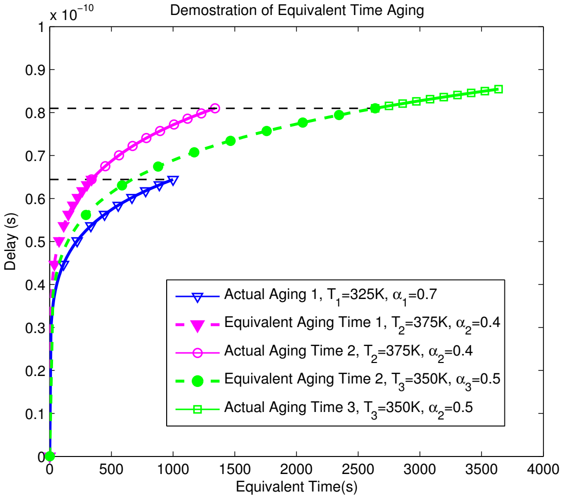

Fig. 3. Demo of equivalent aging time. A NAND2 gate experiences three different conditions consecutively. Each working condition lasts for 1000s.

C. Aging-Aware SSTA and Pruning Algorithm for Repeated Analysis

Once the aging effect of a gate i is computed, the delay of each gate is expressed in PCE form (Equation 9) of its underlying process parameter $\vec{\xi}_i$. To handle the correlation between the process parameters $\{\vec{\xi}_i\}_1^q$ from q different gates, principle component analysis [2] is carried out to extract a set of independent variables $\vec{z}$ according to the correlation matrix of $\{\vec{\xi}_i\}_1^q$. And each set of process parameters $\vec{\xi}_i$ is represented by a linear combination of $\vec{z}$. After this transformation, a PCE delay model based timing analysis method [5] is adopted to compute the arrival time (AT) and required time (RT) of each node on the graph. The arrival time and required time are also represented in the PCE form.

In practice, the circuit needs to be tested under different combinations of working conditions and it is time-consuming and thus undesirable if aging-ware SSTA has to be re-run each time the working condition of the circuit changes. Inspired by the work of [24], we propose a fast pruning algorithm which prunes redundant timing nodes and edges and thus improves the subsequent repeated analysis.

It is observed in [24] that the distributions of the arrival time at the different inputs of a node may be distantly separated. The statistical maximum of them, which is the distribution of arrival time at the output of the node, will also be separated from some of the inputs. Such input can be identified using the following equation:

$$\mu_{out} - \mu_{in} \geq \sigma_{in} + \sigma_{out} \tag{15}$$

where μ and σ are the means and variances of the input and output arrival times. Inputs of this type do not contribute to the timing statistics of the later stages and therefore the incident edges corresponding to this type of inputs can be removed from the timing graph of the circuit. Note that as the circuit ages, it is possible that inputs not contributing to circuit performance at time zero become relevant due to the NBTI aging effect. However, as studied in [21] and [20], the maximum delay degradation along a timing path is limited by 20%. Considering this factor, we add a safe margin to the edge pruning condition to prevent the removal of any potentially important edge due to aging effect:

$$\mu_{out} - \mu_{in}(1+\epsilon) \geq (\sigma_{in} + \sigma_{out}) \times (1+\epsilon) \tag{16}$$

where ϵ is set to be 20%.

Furthermore, a node with all its fan-out edges pruned can also be removed from the timing graph, along with its fan-in edges. In summary, the pruning algorithm consists of two phases. Firstly, it does a forward topological search to prune unimportant timing edges using Equation (16). Next, it carries out a backward topological search to prune nodes with no remaining fan-out edges, and their fan-in edges. The running time of the pruning algorithm is dominated by SSTA in the first phase of the algorithm. The payoff of this algorithm is that the aging-aware SSTA under different working conditions can be sped up on the pruned graph. In addition, optimization needs only to consider the pruned graph.

V. Criticality-Sensitivity Based Element Impact Analysis

Statistical analysis fulfills two purposes: first, it calculates the reliability after several years' aging; second, when the target reliability is not achieved, it provides guidance for circuit optimization. In this section, we propose a criticality-sensitivity based element impact analysis, which can be embedded in an iterative optimization framework to improve reliability of the circuit after years of aging. The whole flow is shown in the lower right part of Figure 1. Our aging optimization procedure is similar to TILOS [9] and the work in [10] in the sense that it chooses a group of most *effective* gates for sizing at each iteration. However, we propose a criticality-sensitivity based analysis to measure, in the probability space, the impact of sizing this gate on the reduction of aging effect of the whole circuit. At each iteration of the optimization flow, the aging-aware statistical timing analysis proposed in Section VI is used to update the aging effect and timing information of the circuit. Then, criticality-sensitivity based element impact analysis discussed below is carried out to select a group of gates Φ, whose change affects the timing of the circuit most effectively. At the end of the iteration, gates in Φ are optimized to improve the circuit yield under process variation and aging effect. The optimization iteration repeats until it meets the expected reliability or violates the constraints of area or power.

Because of circuit structure, the effects of updating different elements are different and interdependent. Updating one element in each iteration solves the interdependence problem but is expensive. It is also unnecessary to update an element on a non-critical path. Therefore, it is important to select a small group of most effective elements to optimize in each iteration. As a study, we focus on gate sizing where the designed width μ_1 of each gate will be selected to optimize the circuit reliability at the end of a given aging period. Notice that the fabricated actual width ξ_1 of a gate is still a random variable, but its mean value and variance are decided by μ_1.

The effect of sizing gate i on the reliability y of the whole circuit, which is defined as the circuit yield after a given period of aging, is just the reliability gradient over μ_1, which can be expressed as

$$\frac{\partial y}{\partial \mu_1} = \frac{\partial y}{\partial A_i} \cdot \frac{\partial A_i}{\partial d_i} \cdot \frac{\partial d_i}{\partial \xi_1} \cdot \frac{\partial \xi_1}{\partial \mu_1} \tag{17}$$

where d_i is the gate delay, and A_i is the slack given by

$$A_i(\vec{\xi}) = AT_{i,in}(\vec{\xi}) + d_i(\vec{\xi}) + RT_{i,out}(\vec{\xi}) \tag{18}$$

where $AT_{i,in}$ and $RT_{i,out}$ are the arrival time at the input and the required time at the output of gate i. From the above relationship, we know $\partial A_i/\partial d_i = 1$. The product of the last two terms in Equation (17) gives the **sensitivity** $\partial d_i/\partial \mu_1$, which is the derivative of the gate delay with respect to the designed width. Since gate delay is a distribution, the sensitivity, as its derivative over a constant variable, is also a distribution. We will use its mean value to guide the optimization. For this, we have

$$E(\partial d_i/\partial \mu_1) = E(\frac{\partial d_i}{\partial \xi_1}) \cdot E(\frac{\partial \xi_1}{\partial \mu_1}) \tag{19}$$

where the first term is a constant and can be obtained from the coefficient of the ξ_1 term in Equation (9), and the second term is 1 if μ_1 gives the mean of ξ_1.

The term $\partial y/\partial A_i$ captures the complex relationship between gate slack and circuit reliability. When the slack is positive, it will be zero. Even when the slack is negative, if the gate is not on all critical paths, the term will still be zero. Furthermore, with process variation, A_i becomes a distribution instead of a deterministic figure. Therefore, we use criticality [25] to evaluate how important a gate is to improving the reliability. Considering the aging effect, the **criticality** c_i is defined as the probability that gate i lies on the critical path of the circuit after a period of aging time, due to the process variation. To compute the criticality of each gate in the circuit, we adopt the cutset-based method proposed in [25], [17]. Since our timing distribution is expressed in PCE form instead of first-order canonical form, the tightness probability, a basic building block of the criticality computation algorithm, is computed using either numerical integration [8] or the APEX method [16].

After the criticality c_i and sensitivity s_i of every gate in the circuit are computed, we rank the gates according to the product $c_i \times s_i$. Although $c_i \times s_i$ is not the same as $\partial y/\partial \mu_1$, it provides an approximate ranking of the gates by their impact on the improvement of the circuit reliability after aging. Using this criterion, a group of n highest-ranking gates are chosen for optimization. After optimization, the aging-aware statistical timing analysis is carried out to update the aging effect, timing, and reliability of the circuit. Notice that in the analysis, not all the gates in the circuit need to be updated. Only the fan-in gates of the n modified gates and those on their fan-out cones need to be reexamined.

VI. Experimental Results

The proposed statistical reliability analysis framework is implemented in C++, including a modeling engine, a timing analysis engine, and a criticality-sensitivity based element impact analysis engine. The effectiveness of the impact analysis is tested in an iterative optimization approach. Circuits from ISCAS85 are used as the testbench. The experiments are run on a 64-bit Linux server with 3.0GHz Xeon CPU and 2G memory.

A. Verification of the Gate Aging Model

This section evaluates the proposed statistical gate aging model shown in Equation 8. Various types of logic gates, such as BUF, NAND, and NOR, are considered with 65nm PTM [1] technology. For each type of gate, the channel width, gate length, and threshold voltage of PMOS and NMOS are modeled using Gaussian random variables. The variance of each random parameter is set as 10% of its mean. The reference aging model is characterized once using the proposed modeling method under normal working condition with $T = 325K$ and $\alpha = 0.5$. To test the accuracy of our aging model in different working conditions, the aging effect of the gates under a different working condition is computed by scaling from the reference model using Equation 10. We select $T = 350K$ and $\alpha = 0.75$ as the testing condition, which achieves the maximum temperature difference against the reference condition allowed in our model, as is discussed in Subsection IV-A. The aging effects of each gate are compared against the results of corresponding 5000-point Monte-Carlo simulation. In each of the Monte-Carlo instances, the random ΔV_{th} is computed using long-term model in Equation 1 according to the sampled process parameters.

The relative error results against Monte-Carlo are given in Table I. The aging effect is modeled accurately at the reference working condition. As the working condition deviates from the reference one, the modeling error begins to grow, which is not unexpected since the modeling error stems not only from the coefficient regression but also from the fact that the scaling relationship in α and T does not exactly satisfy the long-term model. Still, the aging model has sufficient accuracy to be used in the aging-aware statistical analysis framework under allowed fluctuation of working conditions.

TABLE I
Error of gate aging model against Monte-Carlo

Gate	Ref.		Scaled	
	μ	σ	μ	σ
NOT	0.01%	0.15%	2.45%	4.87%
BUF	0.02%	1.12%	1.02%	2.53%
NAND2	0.05%	0.86%	2.19%	5.32%
NOR2	0.11%	1.38%	2.15%	5.34%
XOR2	0.05%	2.67%	2.78%	4.11%
NAND4	0.06%	2.42%	3.98%	4.96%
NOR4	0.21%	1.83%	2.87%	4.25%

B. Aging Effects of Circuit Delay Distribution

Next, using the proposed aging-aware statistical timing analysis method, We characterize circuit delay distributions under aging effects. The proposed analysis method uses the SSTA method which has been validated against Monte-Carlo simulations in the past work [5]. To evaluate temperature-dependent aging effect, we constructed a run-time temperature profile gathered from a computer server. The chip temperature varies from 36°C (low workload) to 75°C (high workload). The temperature profile is then repeated in a 5-year time span. In addition, during the first two years, the ratio between high workload and low workload is set to 0.5, and for the next three years, the ratio is increased to 2. The aging history for ISCAS85 circuits under the changing temperature profile is computed using the proposed equivalent aging time technique. The delay distributions of each circuit are computed at years 0, 2 and 5. The evolution of the means and the variances at years 2 and 5 are shown with respect to year zero in Figure 4.

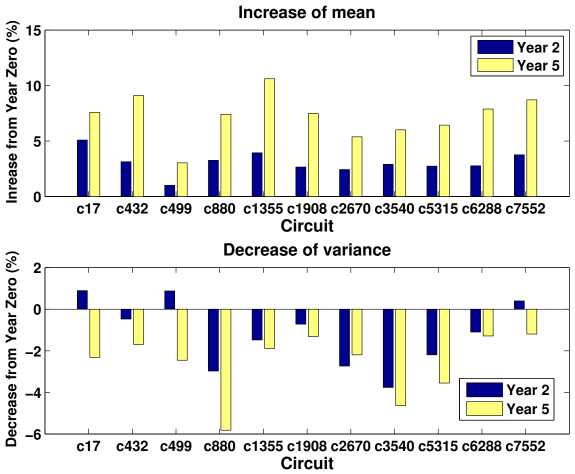

Fig. 4. Degradation of circuit delay distribution due to aging effect.

These results demonstrate that, due to process variation and NBTI aging effect, mean delay increases while delay variance decreases in general, as was first discovered by Wang et al [22]. However, in some circuits, such as c17, c449 and c7552, delay variance increases over time. This is due to the fact that, unlike the single path case [22], the distributions of the paths converged at a gate inputs may approach or depart from each other, resulting in a possible increase of circuit delay variance.

C. Pruning Effects

We next evaluate the effectiveness of the proposed aging-aware pruning algorithm (Section IV-C). We choose $\epsilon = 20\%$ as the delay degradation of the paths in the circuit will not exceed 20% under normal operation, which is also suggested by Figure 4. Table II lists the pruning statistics, accuracy and speedup against the results of using SSTA without pruning. With the maximum difference of mean and variance under 1% and 4%, the pruning algorithm introduces little error for the subsequent timing analysis. From Table II, it is observed that the analysis speedup is proportional to the number of pruned edges and nodes, which depends on the structure of the circuit and the different aging effect of each gate in the circuit. For ISCAS85 benchmarks, speedups from 10% to 70% are achieved. However, if a circuit is well-balanced, that is, the arrival time distributions at the inputs of every node are close to each other, the pruning algorithm is not effective, as shown in the result of c7552.

D. Effectiveness of Element Impact Analysis

Finally, we designed and implemented an iterative gate sizing optimization framework to test the effectiveness of the proposed criticality-sensitivity based element impact analysis (Section V). The four largest benchmarks shown in Table II are used in this experiment. For each benchmark, we consider the high workload condition with temperature of 75°C. The optimization flow is set to target a 5-year time span. During each optimization iteration, the proposed aging-aware timing analysis is used to estimate the circuit timing information. Gates are then ranked by the product of criticality and sensitivity in a non-increasing order. The first Φ gates are then chosen,

TABLE II
PRUNED SSTA WITH AGING EFFECT

Circuit	Pruned	Pruned	Error		Speedup
	Nodes	Edges	μ	σ	
c17	33.3%	35.7%	0.25%	1.48%	1.29X
c432	33.5%	40.8%	0.04%	2.16%	1.24X
c499	22.4%	43.6%	0.06%	2.29%	1.31X
c880	50.6%	58.6%	0.20%	3.21%	1.75X
c1355	19.2%	30.6%	0.03%	2.90%	1.22X
c1908	36.4%	41.4%	0.02%	2.52%	1.46X
c2670	47.4%	51.3%	0.20%	1.98%	1.76X
c3540	42.3%	49.9%	0.08%	3.74%	1.52X
c5315	10.1%	17.3%	0.03%	1.87%	1.06X
c6288	11.9%	22.0%	0.08%	1.31%	1.10X
c7552	6.31%	5.45%	0.02%	2.60%	1.01X

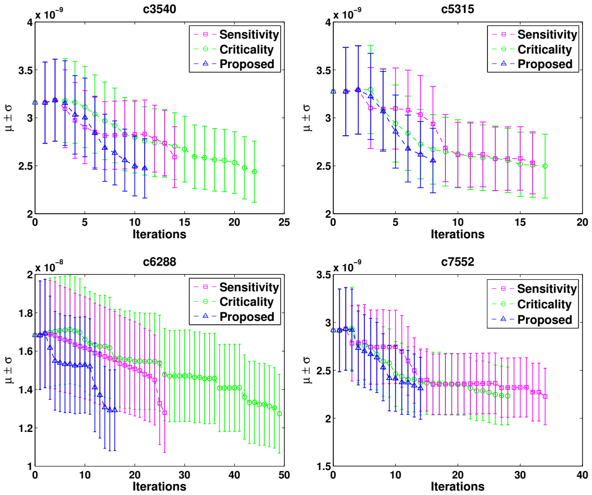

Fig. 5. Error bars of the mean and variance during optimization.

and the size of each gate is increased by ϵ. $\Phi = 50$ provides the best performance-time trade-off in our experiments. This procedure is repeated until the $\mu + \sigma$ of the circuit delay is improved by 25%. For comparison purposes, we also consider another two gate impact analysis methods which rank the gates using either sensitivity or criticality alone. The traces of the mean and variance of the delays of these circuits after each iteration are plotted in Figure 5. It can be observed that the progress of the sensitivity-guided method is unstable while that of the criticality-guided method is not fast enough. On the other hand, the progress of the proposed method using the product of sensitivity and criticality is both stable and fast, showing the effectiveness of criticality-sensitivity analysis.

VII. CONCLUSION

In this work, we have studied the impact of process variation and run-time aging effect on IC lifetime reliability. This work yields a comprehensive IC reliability analysis framework, which considers the joint impact of NBTI aging effect and parameter variations. Techniques are proposed to optimize the accuracy and efficiency of IC reliability analysis, including scalable NBTI aging model, equivalent aging time and aging-aware timing graph pruning. A novel criticality-sensitivity based analysis method is proposed to allow rapid estimation of the impact individual circuit element on the overall circuit lifetime reliability. Leveraging the proposed analysis framework, we have designed and implemented a statistical reliability optimization flow for nanometer-scale IC design.

ACKNOWLEDGMENT

This research is supported partially by NSFC research project 60676018 and 60806013, China National Basic Research Program under the grant 2005CB321701, China National Major Science and Technology special project 2008ZX01035-001-06 during the 11th five-year plan period, the doctoral program foundation of Ministry of Education of China under 200802460068, the International Science and Technology Cooperation program foundation of Shanghai under 08510700100, the program for Outstanding Academic Leader of Shanghai, NSF under CCF-0811270, and SRC under awards 2007-HJ-1593 and 2007-TJ-1589.

REFERENCES

[1] *Predictive technology model, http://www.eas.asu.edu/ptm.*

[2] A. Agarawal, D. Blaauw, and V. Zolotov. Statistical timing analysis for intro-die process variations with spatial correlations. In *ICCAD*, 2003.

[3] M. Alam, K. Kang, B. C. Paul, and K. Roy. Reliability- and process-variation aware design of VLSI design. In *IPFA*, India, 2007.

[4] S. Basu and R. Vemuri. Process variation and nbti tolerant standard cells to improve parametric yield and lifetime of ICs. In *ISVLSI*, 2007.

[5] S. Bhardwaj, P. Ghanta, and S. Vrudula. A framwork for statistical timing analysis using non-linear delay and slew models. In *ICCAD*, Nov. 2006.

[6] S. Bhardwaj, W. Wang, R. Vattikonda, Y. Cao, and S. Vrudhula. Predictive modeling of the NBTI effect for reliable design. In *CICC*, pages 189–192, Sep. 2006.

[7] S. Chakravarithi, A. T. Krishman, V. Reddy, C. F. Machala, and S. Krishnan. A comprehensive framework for predictive modeling of negative bias temperature instability. In *Annual International Reliabilty Physics Symposium*, Phoenix, 2004.

[8] H. Chang, V. Zolotov, S. Narayan, and C. Visweswariah. Parameterized block-based statistical timing analysis with non-gaussian parameters, nonlinear delay functions. In *DAC*, 2005.

[9] J. P. Fishburn and A. E. Dunlop. TILOS: a posynomial programming approach to transistor sizing. In *ICCAD*, 1985.

[10] M. R. Guthaus, N. Venkateswarann, C. Visweswariah, and V. Zolotov. Gate sizing using incremental parameterized statistical timing analysis.

[11] K. Kang, S. P. Park, K. Roy, and M. A. Alam. Estimation of statistical variation in temporal NBTI degradation and its impact on lifetime circuit performace. In *ICCAD*, 2007.

[12] H. Kufluoglu and M. Alam. A generailized reaction-diffusion model with explicit $H - H_2$ dynamics for negative-bias temperature-instability (NBTI) degradation. *IEEE Trans. on Electron Devices*, 54(5):1101–1107, May 2007.

[13] S. Kumar, C. H. Kim, and S. Sapatnekar. An analytical model for negative bias temperature instability. In *ICCAD*, Nov. 2006.

[14] S. Kumar, C. H. Kim, and S. Sapatnekar. NBTI-aware synthesis of digital circuits. In *DAC*, June 2007.

[15] S. Kumar, J. Li, C. Talarico, and J. Wang. A probabilistic collocation method based statistical gate delay model considering process variations and multiple input switching. In *DATE*, 2005.

[16] X. Li. Asymptotic probability extraction for non-normal distribution of circuit. In *ICCAD*, 2004.

[17] D. Mogal, H.Qian, S. Sapatnekar, and K. Bazargan. Clustering based prining for statistical criticality computation under process variation. In *ICCAD*, 2007.

[18] T. Sakurai and A. R. Newton. Alpha-power law mosfet model and its application to cmos invert delay and other formulars. *IEEE Journal of Solid-State Circuits*, 25(2):584–594, Apr. 1990.

[19] D. Shroder and J. A. Babcock. Negative bias temperature instability: Road to cross in deep submicron silicon semiconductor manufacturing. *Journal of Applied Physics*, 94(1), July 2003.

[20] W. Wang, Z. Wei, S. Yang, and Y. Cao. An efficient method to identify critical gates under circuit aging. In *ICCAD*, 2007.

[21] W. Wang, S. Yang, S. Bhaedwaj, R. Vattikonda, S. Vrudhula, F. Liu, and Y. Cao. The impact of NBTI on the performance of combinational and sequential circuits. In *DAC*, June 2007.

[22] W.-P. Wang, V. Reddy, B. Yang, V. Balakrishnan, S. Krishnan, and Y. Cao. Statistical predction of circuit aging under process variation. In *CICC*, 2008.

[23] W.-P. Wang, S.-Q. Yang, and Y. Cao. Node criticality computation for circuit timing analysis and optimization under NBTI effect. In *ISQED*, 2008.

[24] Y. Wang, X. Zeng, J. Tao, H. Zhu, X. Luo, C. Yan, and W. Cai. Adaptive stochastic collocation method (ASCM) for parameterized statistical timing analysis with quadratic delay model. In *ISQED*, 2008.

[25] J. Xiong, V. Zolotov, C. Visweswariah, and N. Venkateswara. Criticality compuatation in parameterized statistical timing. In *DAC*, 2006.

[26] B. Zhang and M. Orshansky. Modeling of nbti-induced pmos degradation under arbitrary dynamic temperature variation. In *ISQED*, 2008.