Scalable Methods for Analyzing the Circuit Failure Probability Due to Gate Oxide Breakdown

Jianxin Fang, and Sachin S. Sapatnekar, *Fellow, IEEE*

***Abstract*—Gate oxide breakdown is an important reliability issue that has been widely studied at the individual transistor level, but has seen very little work at the circuit level. We first develop an analytic closed-form model for the failure probability (FP) of a large digital circuit due to this phenomenon. The new approach accounts for the fact that not every breakdown leads to circuit failure, and shows a 4.8–6.2× relaxation of the predicted lifetime with respect to the pessimistic area-scaling method for nominal process parameters. Next, we extend the failure analysis to include the effect of process variations, and derive that the circuit failure probability at a specified time instant has a lognormal distribution due to process variations. Circuits with variations show 19–24% lifetime degradation against nominal analysis and 4.7–5.9× lifetime relaxation against area-scaling method under variations. Both parts of our work are verified by extensive simulations and proved to be effective, accurate and scalable.**

***Index Terms*—Oxide Breakdown, Failure Analysis, Circuit Reliability, Process Variation**

I. Introduction

Of late, reliability issues have become an increasingly important concern in CMOS VLSI circuits. Oxide breakdown refers to the phenomenon where defects are generated in the SiO$_2$ gate oxide under the continued stress of normal operation over a long period. Eventually, the oxide becomes conductive when a critical defect density is reached at a certain location in the oxide. With device scaling, as electric fields across the gate oxide have increased as supply voltages have scaled down more slowly than the oxide thickness, transistors have become more susceptible to oxide breakdown.

At the device level, the mechanisms and modeling of oxide breakdown have been studied for several decades, yielding a large number of publications, as surveyed in [1]. Various empirical and analytical models, including percolation models [2], [3] have been proposed for this phenomenon. The time-to-breakdown characteristic for a MOS transistor is typically modeled as a Weibull random variable, and characterized by accelerated experiments, in which MOS transistors or capacitors are subjected to high voltage stress at the gate terminal, with both the source and drain terminals grounded until breakdown is detected [4], [5].

The effect of a breakdown is to provide a path for current to flow from the gate to the channel. The terms *hard breakdown* (HBD) and *soft breakdown* (SBD) are widely used to describe the severity of oxide breakdown occurrences. Functional failures, which are the focus of this work, can only be caused by HBDs (although, as we will show, not every HBD causes a functional failure). Unlike in analog or memory circuits where SBDs can provoke circuit failure, SBDs in digital logic circuit can only cause parametric variations but not functional failures [1], [6], [7], therefore they are not considered in this work. Through the rest of this paper, the term “circuit failure” implies a functional failure in digital logic circuits.

It is believed that there is no substantial difference between the physical origins of the HBD and SBD modes [8], and they are generally distinguished by the resistance of the breakdown path and the consequence to the devices. An HBD is a low-resistance breakdown that can cause significant current to flow through the gate, while an SBD has a higher resistance, and lower breakdown current through the gate [1]. A quantitative comparison of these two modes is presented in [9], and the concept of HBD and SBD has been verified for technologies down to 40nm [10].

At the circuit level, the traditional failure prediction method for a large circuit uses area-scaling, extrapolated from single device characterization [1]. The idea is based on the weakest-link assumption, that the failure of any individual device will cause the failure of the whole chip. Recently, new approaches have been proposed to improve the prediction accuracy by empirical calibration using real circuit test data [11], or by considering the variation of gate-oxide thickness [12]. The former is empirical and hard to generalize, while the latter does not consider the effect of breakdown location. Moreover, all existing methods circuit-level methods assume that (a) the transistors in the circuit are *always* under stress, and (b) any transistor breakdown *always* leads to a circuit failure. These assumptions are not always true, as discussed in Section II-A.

Precise analysis or measured results on several small circuits have been published, based on the post-breakdown behavior models: for a 41-stage ring oscillator in [13], a 6T SRAM cell in [6], and current mirrors and RS latches in [7]. These methods, using either complex analysis models or are based on measurements, and cannot easily be extended to general large-scale digital circuits in a computationally scalable manner.

On the other hand, the probability of circuit failure is significantly affected by on-chip process variations. Recent work [12] proposed a statistical approach for full-chip oxide reliability analysis considering process variation of T_{ox}; however, this work did not present a path to determining the full distribution of the reliability function or statistics such as its variance. Subsequent work in [14] improved upon this by presenting a post-silicon analysis and mitigation method involving on-chip sensors and voltage tuning. The major drawback of these variation-aware approaches for circuit-level oxide reliability

J. Fang and S. S. Sapatnekar are with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455, USA (e-mail: {fang0116,sachin}@umn.edu).

analysis is that they are all based on the simple notion of area-scaling, which is too pessimistic for circuit lifetime prediction.

The contribution of our work is twofold. First, we develop a scalable method for analyzing the failure probability (FP) of large digital circuits, while realistically considering the circuit environment that leads to stress and oxide breakdown. To achieve this goal, at the *transistor* level, we revise the Weibull time-to-breakdown model to incorporate the actual stress modes of transistors. We propose a new piecewise linear/log-linear resistor model for post-breakdown behavior of transistors as a function of the breakdown location within the transistor, in accordance with device-level experimental data in [9]. At the logic *cell* level, we devise a procedure for performing precise FP analysis for standard cell based digital circuits, and present an effective library characterization scheme. In particular, we demonstrate the circuits have inherent resilience to failure due to gate oxide breakdown, and we use this information to build a characterization methodology and analysis method that provides more correct FP computations than the area-scaling model. At the *circuit* level, we derive a closed-form expression for the FP of large digital logic circuits, based on the above characterization of the post-breakdown circuit operation. This analysis leads to the conclusion that area-scaling estimates are unduly pessimistic.

Second, we explore the effects of process variations on the FP, and find that the predicted FP under nominal condition is significantly affected by variations. We then extend the nominal case FP analysis to include the effect of process variations, and show that this still provide substantially better improvements in the predicted lifetime over the conventional area-scaling model. The transistor-level model and cell-level analysis are updated for process variations, and it is derived that the circuit failure probability at a specified time instant has a lognormal distribution due to process variations, and this distribution expands as the process variations and spatial correlation increase. Both parts of our work are verified by extensive simulations and results prove the proposed methods are effective, accurate and scalable.

We begin with an analysis of the nominal case. Section II presents an overview of transistor-level breakdown models, the post-breakdown behavior, and the value of the breakdown resistance, and introduces our empirical model. Next, Section III develops a method for cell-level FP computation. This is applied to circuit-level calculations in Section IV, where we derive a closed-form formula predicting the circuit-level FP. The theory for the nominal case is extended to variation-aware oxide reliability analysis in Section V. Finally, Section VI presents simulation results to validate the proposed methods, and we conclude in Section VII.

II. Transistor-Level Models

In this section, we discuss models for the time-to-breakdown and the post-breakdown behavior of a transistor. Sections II-A and II-B largely overview existing models, while Section II-C presents our new simple quantitative model for breakdown resistance that can be calibrated from experimental data.

Our discussion is guided by two observations:

- As shown in [1], only hard breakdowns cause serious device degradations.
- As demonstrated in [5], the occurrence of hard breakdown is very prevalent in NMOS transistors but relatively rare in PMOS devices.

Therefore, we only consider NMOS hard breakdown in this work. However, the framework presented here can easily be extended to the cases where these two assumptions are relaxed.

Furthermore, we assume that a transistor will be affected by at most one HBD. This assumption is reasonable, and is similar in spirit to the single stuck-at fault assumption in the test arena: due to the statistical and infrequent nature of breakdown events, the probability of more than one independent breakdown striking the same transistor is very low.

A. Time-to-Breakdown

The transistor time-to-breakdown, T_{BD}, is widely modeled as a Weibull distribution with an area-scaling formula [4]. The breakdown probability of device i, with area a_i, at time t is

$$\Pr_{\mathrm{BD}}^{(i)}(t) = 1 - \exp\left(-\left(\frac{t}{\alpha}\right)^{\beta} a_i\right), \tag{1}$$

where α is the characteristic time corresponding to 63.2% of breakdown probability for the unit-size device with area $a_i = 1$, and β is the Weibull shape factor, also known as the Weibull slope. A common representation of a Weibull distribution is on the so-called *Weibull scale*, under the transform

$$W = \ln(-\ln(1 - \Pr)) = \beta \ln(t/\alpha) + \ln(a_i) \tag{2}$$

In other words, if we plot W as a function of $\ln(t)$, the result is a straight line with slope β.

The Weibull parameters α and β in are usually characterized in experiments, as described in [4], [9], where the gate oxide of the transistor is placed in inversion mode and subjected to a constant voltage stress. However, this experimental scenario is not an accurate representation of the way in which transistors function in real circuits, where the logic states at the transistor terminals change over time, with six possible static stress modes for a NMOS transistor, as shown in Fig. 1[1].

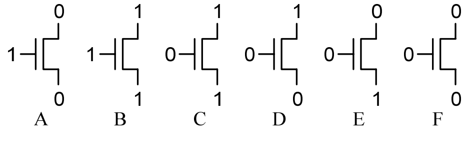

Fig. 1. Stress modes for NMOS transistors.

An HBD occurs in the case of NMOS stressed in inversion, while an NMOS in accumulation almost always experiences SBD [5]. In Fig. 1, Mode A corresponds to inversion, and Modes C, D and E to accumulation, while B and F do not impose a field that stresses the gate oxide. Thus, only the portion of time when the transistor is stressed in Mode A is effective in causing HBDs in a device, and potential circuit

[1] The other two combinations, with the gate at logic 1 and the source and drain at different voltages, are transient modes, not relevant for analyzing long-term stress.

failure. We introduce the stress coefficient, γ_i, for device i to capture the proportion of this effective stress time, and reformulate (1) as

$$\Pr_{\mathrm{BD}}^{(i)}(t) = 1 - \exp\left(-\left(\frac{\gamma_i t}{\alpha}\right)^{\beta} a_i\right) \tag{3}$$

where $(\gamma_i t)$ represents the effective stress period after time t of circuit operation. The stress coefficient γ_i is the probability of Mode A, and can be represented by the joint probability mass function (jpmf) that the (gate, source, drain) terminals of transistor i have the logic pattern $(1, 0, 0)$. This can be calculated using the signal probability (SP) of each node, and maps on to a well-studied problem in CAD. These probabilities may be computed, for example, more approximately by using topological methods that assume independence [15], or using more computational methods that explicitly capture correlations, such as Monte Carlo approaches [16].

B. Post-Breakdown Behavior

Fig. 2(a) shows a two-dimensional schematic that displays the idea of oxide breakdown in a MOS transistor. The channel length is denoted by L, and the source/drain extensions are of length L_{ext}. The distance from the source is denoted by x, and the breakdown is assumed to be located at x_{BD}.

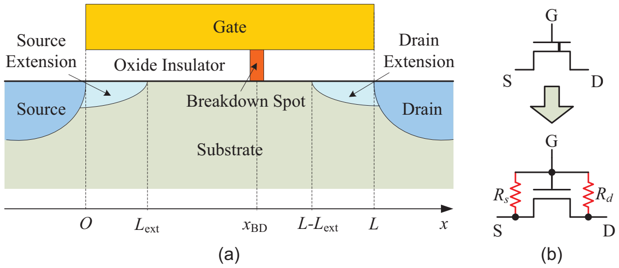

Fig. 2. (a) Schematic of oxide breakdown in a transistor. (b) Resistor model for post-breakdown behavior.

Various modeling approaches for post-breakdown analysis at the transistor- or cell-level have been proposed in the literature. Several approaches have proposed models for SBDs, e.g., [17], [18], but these result in parametric failures rather than the functional failures that this work studies. The work in [19] suggests a complex physical model that reduces to a simple resistor model when the breakdown location is near the source or drain. As summarized in [1], independent experiments have reported that HBDs show a roughly linear (ohmic) I-V characteristic. Based on this, we use a simpler linear resistor model, similar to that in [20], [21], for post-breakdown behavior analysis. A MOS transistor that has undergone oxide breakdown is replaced with a healthy clone and two resistors, R_s and R_d, as shown in Fig. 2(b). The values of these two resistors are dependent on the breakdown location, x_{BD}.

In characterizing the values of these resistances, it is important to lay down some requirements that they must fulfill. Fig. 3(a) shows the experimental measurement value of the effective breakdown resistance, R_{BD}, for hard breakdowns as a function of x_{BD}, where both the source and drain nodes of the transistor are grounded, and R_{BD} is measured between the gate node and the ground [9]. The data points in this figure correspond to measurements, while the solid line is based on a detailed device simulation. Further experimental data in [9] (not shown here), demonstrate that over a range of channel lengths, the nature of the variation of R_{BD} with x_{BD} shows the same trend as in the figure. Specifically, the observations drawn from [9] are that:

- R_{BD} is smaller when the breakdown occurs in the source or drain extension regions, and is larger for x_{BD} in the channel.
- R_{BD} decreases exponentially (note the log scale on the y-axis) when x_{BD} approaches either end of the channel, while it does not vary significantly with x_{BD} in the center of the channel.
- The statistics of the breakdown location, x_{BD}, show a uniform distribution over the length of the channel.

In advanced high-k technology, [22] indicated that breakdowns are more likely to happen in the grain boundary (GB) sites, which also have uniform distribution in the dielectric layer.

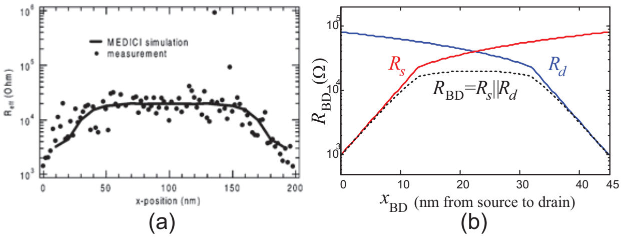

Fig. 3. (a) The effective breakdown resistance as a function of the breakdown location [9]. (b) Modeling of the breakdown resistors.

C. Modeling the Breakdown Resistors

While the structure of the breakdown resistor model using R_s and R_d in Fig. 2(b) is not fundamentally new, there has been less work on deriving a model that relates the breakdown resistance with x_{BD}. The only known work is an equivalent circuit model in [19], but it requires a complex characterization process; moreover, the nonlinearity of the model makes its evaluation in a circuit simulator more time-consuming. We derive a much simpler model based on the idea of fitting the result from experiments and simulation which requires very few measurements for characterization.

The form of the model is guided by the breakdown resistance vs. x_{BD} curve in Fig. 3(a). We propose to capture the variation of the breakdown resistance with x_{BD} through a piecewise linear/log-linear model, where R_s $[R_d]$ varies exponentially with x_{BD} in the source [drain] extension region, and linearly in the remainder of the channel:

$$R_s(x) = \begin{cases} ae^{bx}, & 0 \le x \le L_{\text{ext}} \\ kx, & L_{\text{ext}} \le x \le L \end{cases} \tag{4}$$

Due to source-drain symmetry, we obtain $R_d(x) = R_s(L - x)$. When both the source and drain nodes are grounded, $R_{\text{BD}}(x) = R_s(x) \parallel R_d(x)$. The value of R_{BD} is at its minimum, $R_{\text{BD}\,\text{min}}$, at $x = 0$ and $x = L$, and by symmetry, at its maximum, $R_{\text{BD}\,\text{max}}$ at $x = L/2$. This discussion about

R_{BD} is purely for illustration purposes: in our work, we do not directly use R_{BD}, but work with the R_s and R_d models in conjunction with MOS transistor models.

The constants k, a and b are obtained from experiment measurements in [9] by matching a set of boundary conditions. At $x = 0$, the value of R_s dominates the value of R_d, so that $R_{\mathrm{BD}} \simeq R_s(0) = R_{\mathrm{BD\,min}}$. Thus,

$$a = R_{\mathrm{BD\,min}} \tag{5}$$

At $x = L/2$, by symmetry, $R_s = R_d$, implying that $R_{\mathrm{BD}} = R_s(L/2)/2 = R_{\mathrm{BD\,max}}$. Therefore,

$$k = \frac{4R_{\mathrm{BD\,max}}}{L} \tag{6}$$

Finally, to ensure the continuity between the linear and log-linear pieces of the piecewise model, we must ensure that $\lim_{x\to L_{\mathrm{ext}}^-} R_s(x) = \lim_{x\to L_{\mathrm{ext}}^+} R_s(x)$, i.e., $kL_{\mathrm{ext}} = ae^{bL_{\mathrm{ext}}}$. So,

$$b = \frac{1}{L_{\mathrm{ext}}} \ln\left(\frac{4R_{\mathrm{BD\,max}}L_{\mathrm{ext}}}{R_{\mathrm{BD\,min}}L}\right) \tag{7}$$

Four parameters are required to characterize this model: L, L_{ext}, $R_{\mathrm{BD\,min}}$ and $R_{\mathrm{BD\,max}}$. Fig. 3(b) shows an example plot for the parallel combination of R_s and R_d using this model, with the parameters $L = 45$nm, $L_{\mathrm{ext}} = 13$nm, $R_{\mathrm{BD\,max}} = 20\mathrm{k}\Omega$, and $R_{\mathrm{BD\,min}} = 1\mathrm{k}\Omega$[2]. It is easy to see that the results here are well matched to the trend of experimental measurements in Fig. 3(a).

III. Cell-Level Failure Analysis

Our entire technique for digital circuit failure analysis due to gate oxide breakdown is summarized in Figure 4. At the transistor level, the process parameters L and L_{ext}, and HBD resistance measurements $R_{\mathrm{BD\,max}}$ and $R_{\mathrm{BD\,min}}$ are provided as inputs to the method and utilized to characterize our post-breakdown $R_{\mathrm{BD}}(x_{\mathrm{BD}})$ model using (5–7). The values of α and β for the Weibull distribution that characterizes transistor-level failure are also provided as input parameters. At the logic cell level, the driver and load I-V curves of each logic cell in the input cell library are precharacterized and stored into LUTs using Algorithm 1, which will be described in this section. The calculation of cell FP is performed in a circuit-specific context with Algorithm 2, also described later. Finally the circuit FP analysis is performed using the proposed method, using the result (18) presented in Theorem 1.

This section focuses on analyzing the effects of oxide breakdown at the logic cell level. A formula for the FP for each breakdown case is developed, and a library characterization scheme is proposed for standard cell based digital circuits.

A. Breakdown Case Analysis

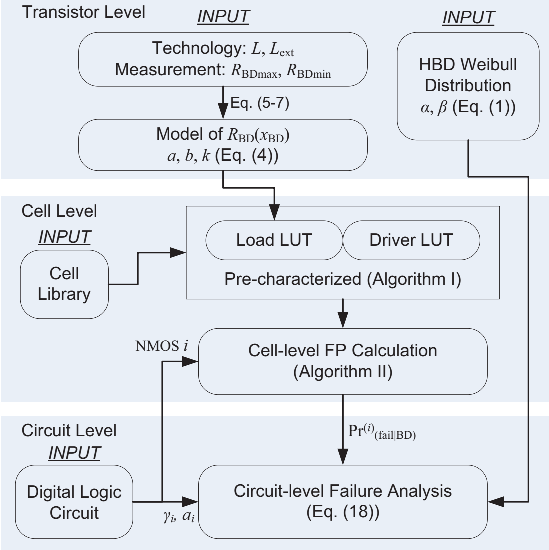
Transistor Level
INPUT
Technology: L, L_{ext}
Measurement: R_{BDmax}, R_{BDmin}
Eq. (5-7)
Model of $R_{\mathrm{BD}}(x_{\mathrm{BD}})$
a, b, k (Eq. (4))
INPUT
HBD Weibull Distribution
α, β (Eq. (1))
Cell Level
INPUT
Cell Library
Load LUT
Driver LUT
Pre-characterized (Algorithm I)
NMOS i
Cell-level FP Calculation (Algorithm II)
$\mathrm{Pr}^{(i)}_{(\mathrm{fail|BD})}$
Circuit Level
INPUT
Digital Logic Circuit
γ_i, a_i
Circuit-level Failure Analysis (Eq. (18))

Fig. 4. Flow chart of digital circuit oxide reliability analysis.

The effect of the gate oxide breakdown in an NMOS transistor is to create current paths from the gate node of the transistor to its source and drain nodes. In CMOS circuits, the gate node of a device is typically connected to the output of another logic cell or latching element, while the source/drain nodes are, by definition, connected to transistors within the same logic cell (or more generally, the same channel-connected component). This implies that while analyzing breakdown at the gate node of a transistor, it is necessary to consider both the logic cell that it belongs to and the preceding logic cell that drives the gate node of the transistor.

Consider a cell n that contains a transistor with oxide breakdown. Let k be the pin of cell n connected to the gate of this transistor, and let m be the logic cell that drives pin k of cell n. Then for any broken down NMOS transistor, we can find the corresponding case index (m, n, k). Fig. 5(a) shows an example of such a breakdown case, using a NAND2 as cell m, a NOR2 as cell n, and $k = 1$. Here we call cell m as the *driver cell* and cell n as the *load cell* of this case.

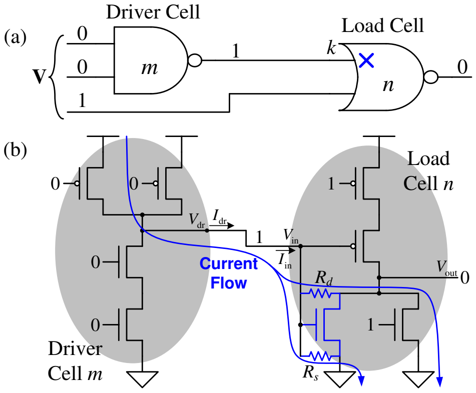

Fig. 5. Cell-level analysis of the breakdown case.

To analyze each breakdown case (m, n, k), we must specify the input vector $\mathbf{V}$ for the free pins of the two cells. The input vector $\mathbf{V}$ is a Boolean vector of dimension $q(m,n) = (\mathrm{Fanin}(m) + \mathrm{Fanin}(n) - 1)$, i.e., $\mathbf{V} \in \mathbb{B}^{q(m,n)}$, where $\mathrm{Fanin}(i), i \in \{m, n\}$ represents the number of input pins of cell i; in Fig. 5(a), q = 3, and we consider the assignment $\mathbf{V} = (0, 0, 1)$. We refer to a breakdown case for a

[2] The values of $R_{\mathrm{BD\,max}}$ and $R_{\mathrm{BD\,min}}$ are input parameters and independent of the analysis approaches. Based on projections from the published literature, their values are taken to be $20\mathrm{k}\Omega$ and $1\mathrm{k}\Omega$, respectively.

specific input vector as $(m, n, k, \mathbf{V})$. Any given $(m, n, k, \mathbf{V})$ combination can be analyzed based on the post-breakdown behavior model discussed in Section II. The transistor-level circuit, using the resistor model, is shown in Fig. 5(b), with the current flow path due to oxide breakdown indicated. The worst case, over all input vectors (it should be noted that q is a small number) for this two-cell structure defines the failure probability, as quantified in the next subsection.

Essentially, Fig. 5(b) shows that the current lost due to the breakdown event has the potential to alter the logic value at the output of cell m or n or both; whether it actually does so or not depends on the strength of the opposing transistor that attempts to preserve the logic value.

B. Calculation of Failure Probabilities (FPs)

The breakdown case in Fig. 5 is analyzed using SPICE DC sweep over x_{BD} with 45nm PTM model [23] and $V_{\text{dd}} = 1.2$V. The output voltages of driver cell m and load cell n, denoted by V_{dr} and V_{out}, as functions of x_{BD}, are shown in Fig. 6. This figure indicates that when breakdown occurs near the source or drain and the breakdown resistor, R_s or R_d, is small, the output voltages of cells m and n may shift away from their nominal values of V_{dd} and 0, respectively. Beyond certain limits, the logic could flip and result in circuit failure.

Note that the results for driver and load cells are asymmetric for the input excitation in Fig. 5, in that the driver cell m shows a failure when the defect lies at either end of the channel, while the failure for the load cell n appears only when the defect lies at the drain end. The difference lies in the case that x_{BD} is small where R_s is very small and R_d is large. In this case the other NMOS in cell n is on and the output voltage is relatively unaffected even in the presence of a breakdown.

We introduce two thresholds, V_H and V_L (in the figure, $V_H = 0.7V_{\text{dd}}, V_L = 0.3V_{\text{dd}}$), so that if the voltage surpasses these thresholds, a failure is deemed to occur. It can be shown that since the variation of the resistance with x_{BD} is monotonic near the drain [source], and since MOS transistors typically have monotonically increasing I-V curves, the output voltages of the impacted logic cells will also change monotonically with x_{BD} near the drain [source]. In other words, the *failure region* on either side of the channel is a continuous interval[3], which is determined by the corresponding crossover point. We define the crossover points to be $x^{\text{dr}}_{\text{fail-s}}, x^{\text{ld}}_{\text{fail-s}}, x^{\text{dr}}_{\text{fail-d}}$, and $x^{\text{ld}}_{\text{fail-d}}$, which refer to the breakdown locations where the corresponding cell output voltages cross the threshold, as illustrated in Fig. 6[4].

This result is not surprising: the breakdown resistance is large in the channel and small in the source/drain extension regions, so that breakdowns in the latter regions are liable to cause logic failures.

We can then obtain the source-side and drain-side *failure probability* (FP) separately for this specific breakdown case and input vector by evaluating the probability of x_{BD} falling within the corresponding failure region. According to [9], [22], the breakdown position is uniformly distributed in the channel, i.e., $x_{\text{BD}} \sim \text{U}[0, L]$. Therefore, these FPs are given by:

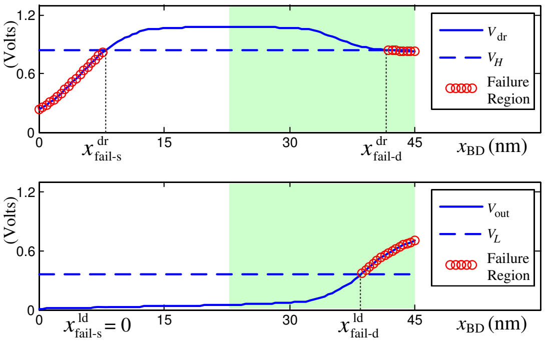

Fig. 6. Cell output voltages under breakdown.

$$\begin{aligned} \text{Pr}^{(m,n,k,\mathbf{V})}_{(\text{fail-s}|\text{BD})} &= \max\left(p^{\text{dr}}_s, p^{\text{ld}}_s\right) \\ \text{Pr}^{(m,n,k,\mathbf{V})}_{(\text{fail-d}|\text{BD})} &= \max\left(p^{\text{dr}}_d, p^{\text{ld}}_d\right) \end{aligned} \tag{8}$$

where, for a given breakdown case $(m, n, k, \mathbf{V})$, the FP components are

$$\begin{aligned} p^{\text{dr}}_s &= \frac{x^{\text{dr}}_{\text{fail-s}}}{L}, \qquad p^{\text{dr}}_d = 1 - \frac{x^{\text{dr}}_{\text{fail-d}}}{L} \\ p^{\text{ld}}_s &= \frac{x^{\text{ld}}_{\text{fail-s}}}{L}, \qquad p^{\text{ld}}_d = 1 - \frac{x^{\text{ld}}_{\text{fail-d}}}{L} \end{aligned} \tag{9}$$

A transistor breakdown with case index (m, n, k) corresponds to a logic failure if such a failure is seen under any input vector $\mathbf{V} \in \mathbb{B}^{q(m,n)}$. This is because once the device-level failure occurs, the circuit is considered to functionally fail if it fails under *any* input vector. Therefore the FP of either side for case (m, n, k) is the worst over all input vectors $\mathbf{V} \in \mathbb{B}^{q(m,n)}$, i.e., the maximum probability among all input vectors. Under the assumption of at most one HBD per transistor, the events of source-side failure and drain-side failure are mutually exclusive, therefore the total FP for case (m, n, k) is the sum of the two sides:

$$\text{Pr}^{(m,n,k)}_{(\text{fail}|\text{BD})} = \max_{\mathbf{V}\in\mathbb{B}^q} \text{Pr}^{(m,n,k,\mathbf{V})}_{(\text{fail-s}|\text{BD})} + \max_{\mathbf{V}\in\mathbb{B}^q} \text{Pr}^{(m,n,k,\mathbf{V})}_{(\text{fail-d}|\text{BD})} \tag{10}$$

Since the logic cells come from a common cell library, $\mathbb{C}$, it is possible to characterize a library over all breakdown cases as a precomputation. For circuit-level failure analysis, as described in Section IV, the precomputed FP results can be retrieved from the characterized library in $O(1)$ time.

C. Cell Library Characterization

The principles behind our cell-level failure analysis procedure have been outlined in the previous two subsections. However, the implementation of this approach involves the analysis of cases $(m, n, k, \mathbf{V})$, and a simple precharacterization would involves a quadratic-complexity enumeration of both driver and load cells from the library. Specifically, the number of SPICE simulations required for this precharacterization, N_{enum}, is computed as:

$$N_{\text{enum}} = N^2_{\text{cell}} \cdot N_{\text{pin}} \cdot 2^{2N_{\text{pin}}-1} \tag{11}$$

Here, N_{cell} stands for the number of cells in the library, and N_{pin} is a bound on the number of fan-ins for a cell; practically,

[3] If the output voltage does not cross the threshold, the failure region may be an empty set, as in the left part of the lower graph of Fig. 6.

[4] If no crossing point exists, the value of the parameter is set to zero at the source end or L at the drain end.

this is a small constant (and this is substantiated on a Nangate library in our experimental results). The number of enumerations, N_{enum}, is the total possible combinations of $(m, n, k, \mathbf{V})$, with $m, n \leq N_{\text{cell}}$, $k \leq N_{\text{pin}}$, and Boolean vector $\mathbf{V}$ has $2^{2N_{\text{pin}}-1}$ combinations. With N_{pin} well bounded ($N_{\text{pin}} \leq 6$ in the Nangate library we used), the case amount $N_{\text{enum}} \propto N_{\text{cell}}^2$ has quadratic complexity with library size, which presents a problem for the cell library characterization process, especially for libraries with a larger number of cells. For example, experiments on a 55-cell Nangate library show that about 1.7 million such enumerations are necessary: clearly, this is a very high cost, even for a one-time precharacterization step.

To overcome this cost without any significant sacrifice in accuracy, we propose a method that improves the scalability of our failure analysis approach. The essence of the idea is that instead of precharacterizing and storing all quadratic combinations, we precharacterize the I-V curves for the library cells and then solve the breakdown cases on the fly. The number of precharacterizations is linear in the number of cells, and the solution can be performed in constant time. Specifically, our library characterization and cell-level FP calculation scheme consists of two stages:

- In the first stage (*precharacterization*), we consider the possibility that each library cell may feature as a driver for another load gate and a load for another driver gate. Accordingly, each cell is characterized to obtain its driver I-V curve (when it acts a driver cell) and its load I-V-x_{BD} curve (when it acts as a load cell) separately, the curves are stored numerically in look up tables (LUTs).
- In the second stage (*FP calculation*), which is performed during the analysis of a specific circuit, the precharacterized curves are used to compute the FP of a specified $(m, n, k, \mathbf{V})$ case from the I-V curves of the driver cell and the load cell using the LUT data.

Fig. 7 shows an example that demonstrates our improved scheme. For the example shown in Fig. 5, Fig. 7(a) plots the precharacterized $I_{\text{dr}}(V_{\text{dr}})$ curve for the driver cell and the precharacterized family of $I_{\text{in}}(V_{\text{in}}, x_{\text{BD}})$ curves (indexed by x_{BD}) for the load cell, and these capture the interaction between the driver and the load cell at the output of the driver. The effect on the output voltage of the load cell is captured by Fig. 7(b), which shows the precharacterized family of curves for $V_{\text{out}}(V_{\text{in}}, x_{\text{BD}})$, indexed by the value of x_{BD}. Note that the load curves are shown for $x_{\text{BD}} \in [L/2, L]$, i.e., the drain side, and in this range, I_{in} and V_{out} are monotonic function of x_{BD}. As we will show later, these curves are adequate to capture the interaction between the driver and the load in any circuit.

Algorithm 1 presents the precharacterization procedure that precomputes these curves. Note that this precharacterization is performed off-line, like standard cell characterization, and must be carried out just once for a given technology. The complexity of this algorithm is linear in the size of the cell library, and the notations used within the algorithm are as follows: for a cell i, I_{dr} and V_{dr} stand for the current and voltage when the output pin of the cell acts as a driver; I_{in} and V_{in} stand for the input current and voltage when the input pin k of the cell acts as a load; and V_{out} stands for the voltage of the output pin of cell i when it acts as a load.

As mentioned earlier, each cell i in the library is characterized separately in its role as a driver and as a load. For the driver characterization, the $I_{\text{dr}}(V_{\text{dr}})$ curve is calculated with sampled values for V_{dr}, for all possible input combinations. Therefore the total number of driver I-V LUTs is $N_{\text{cell}} \cdot 2^{N_{\text{pin}}}$. The load characterization is performed similarly but with an additional enumeration that samples the breakdown location, x_{BD}. The total number of $I_{\text{in}}(V_{\text{in}}, x_{\text{BD}})$ and $V_{\text{out}}(V_{\text{in}}, x_{\text{BD}})$ LUTs corresponding to this is $2 \cdot N_{\text{cell}} \cdot N_{\text{pin}} \cdot 2^{N_{\text{pin}}}$. The storage overhead associated with all driver and load LUTs in the entire library is given by

$$\begin{aligned} \text{Driver} &: \; N_{\text{cell}} \cdot 2^{N_{\text{pin}}} \cdot N_V, \\ \text{Load} &: \; 2 \cdot N_{\text{cell}} \cdot N_{\text{pin}} \cdot 2^{N_{\text{pin}}} \cdot N_{x_{\text{BD}}} \cdot N_V \end{aligned} \tag{12}$$

where N_V stands for the number of V_{dr} and V_{in} samples, and $N_{x_{\text{BD}}}$ stands for the number of x_{BD} samples. This implies that the storage is linear in N_{cell} since the other terms in this expression are bounded by moderate constants in practice.

Algorithm 1 The characterization of cell library for FP calculation.

```
{Driver characterization}
for each cell $i$ in the library do
    for each input vector $\mathbf{V}$ of cell $i$ do
        Calculate $I_{\text{dr}}(V_{\text{dr}})$ for samples of $V_{\text{dr}}$
        Store $I_{\text{dr}}(V_{\text{dr}})$ in driver LUT for cell $i$ input $\mathbf{V}$
    end for
end for
{Load characterization}
for each cell $i$ in the library do
    for each input pin $k$ of cell $i$ do
        for each input vector $\mathbf{V}$ of cell $i$ do
            Calculate $I_{\text{in}}(V_{\text{in}}, x_{\text{BD}})$ and $V_{\text{out}}(V_{\text{in}}, x_{\text{BD}})$ for samples of $V_{\text{in}}$ and $x_{\text{BD}}$
            Store $I_{\text{in}}(V_{\text{in}}, x_{\text{BD}})$ and $V_{\text{out}}(V_{\text{in}}, x_{\text{BD}})$ in load LUT for cell $i$ pin $k$ input $\mathbf{V}$
        end for
    end for
end for
```

Using these precharacterized curves, the second stage, FP calculation, is applied in a circuit-specific context. Given a driver cell and a load cell, the FP calculation step must compute the unknown voltages at the output of the driver and the load. We now demonstrate this calculation for the scenario in Fig. 5, where the correct outputs of the driver and load cell correspond to logic 1 and 0, respectively. For this scenario, the following circuit equations must be solved to determine the unknown voltages:

$$\begin{aligned} I_{\text{dr}}(V_{\text{dr}}) &= I_{\text{in}}(V_{\text{in}}, x_{\text{BD}}) \\ V_{\text{dr}} &= V_{\text{in}} \\ V_{\text{out}} &= V_{\text{out}}(V_{\text{in}}, x_{\text{BD}}) \end{aligned} \tag{13}$$

Consider the problem of solving this for a HBD on the drain side, $x_{\text{BD}} \in [L/2, L]$, affecting the voltage at the driver output, V_{dr}, as illustrated by the failure region on the right of

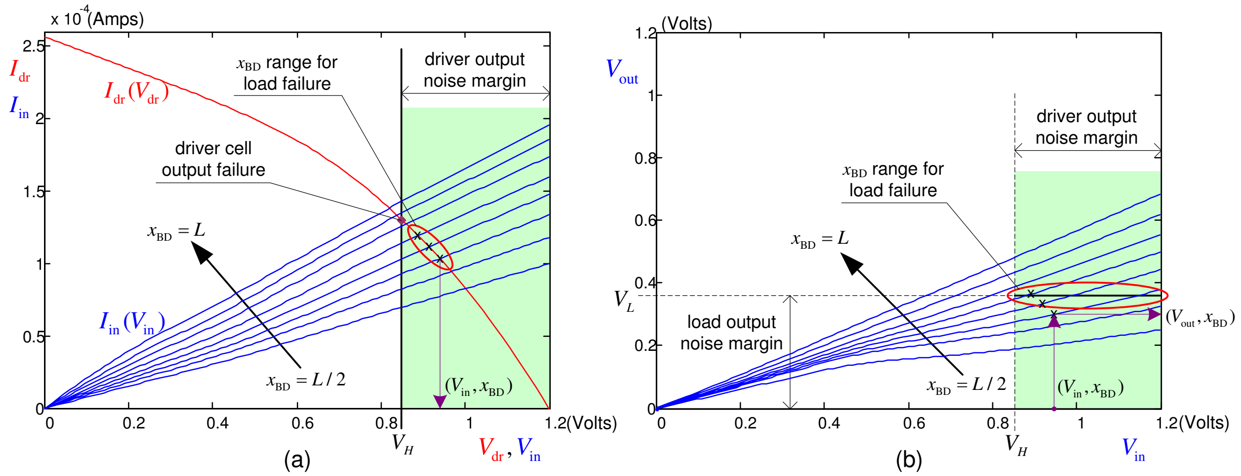

Fig. 7. Demonstration of solving the cell-level FP using I-V curves of the driver and load cells.

the V_{dr} curve in Fig. 6. From (13), $I_{\text{in}}(V_H, x_{\text{BD}}) = I_{\text{dr}}(V_H)$, corresponding to the intersection of two plots and the $V_{\text{dr}} = V_{\text{in}} = V_H$ line in Fig. 7(a). Therefore, for a specific value of V_H, the RHS of this equation can be obtained from the lookup table for the driver side gate. Finding the x_{BD} that solves the equation is then a matter of a reverse lookup on the lookup table for the load side gate.

At any value of V_{in}, since the family of $I_{\text{in}}(V_{\text{in}}, x_{\text{BD}})$ curves increases monotonically with x_{BD}, a failure at $x_{\text{BD}} = x_1$ implies a failure for all $x_{\text{BD}} \geq x_1$, and this solution corresponds to the edge of the failure region, $x^{\text{dr}}_{\text{fail-d}}$, shown in Fig. 6.

Now consider a failure at the load output, V_{out}. Since our goal is to sum up a set of disjoint probabilities, it is important only to consider load output failures that *do not* cause a driver output failure. The procedure consists of two steps:
(1) we consider all intersections in Fig. 7(a) between the I_{dr} and the family of I_{in} curves in the region $V_H \leq V_{\text{in}} \leq V_{\text{dd}}$, and for each of these, we determine the $(V_{\text{in}}, x_{\text{BD}})$ value, and (2) we use the traced $(V_{\text{in}}, x_{\text{BD}})$ values in Fig. 7(b), using the V_{out} LUTs to determine the corresponding value of V_{out}: if this exceeds the threshold, V_L, then we have a failure.

In principle, a drain-side failure that occurs anywhere in the interval $[L/2, L]$ could cause a load output failure. However, we narrow down this range further. The idea is based on the observation that the $V_{\text{out}}(V_{\text{in}}, x_{\text{BD}})$ curves in Fig. 7(b) cross V_L in the interval $V_H \leq V_{\text{in}} \leq V_{\text{dd}}$ only for a specific, typically small, range of x_{BD}. We exploit this idea to improve the efficiency of this procedure, restricting the search in the previous paragraph to this interval of x_{BD}: this is seen to yield considerable computational savings in practice.

To be general, the above idea must be extended to several cases, corresponding to breakdowns at the output of the driver and the load at both possible logic values, due to failures at the drain side and the source side. Thus, we must consider:

$$\begin{aligned}
V_{\text{dr}} &= V^{\text{dr}}_{\text{TH}},\ x_{\text{BD}} \in [0, L/2] \text{ for } x^{\text{dr}}_{\text{fail-s}};\ \text{or} && (14)\\
V_{\text{dr}} &= V^{\text{dr}}_{\text{TH}},\ x_{\text{BD}} \in [L/2, L] \text{ for } x^{\text{dr}}_{\text{fail-d}};\ \text{or} && (15)\\
V_{\text{out}} &= V^{\text{out}}_{\text{TH}},\ x_{\text{BD}} \in [0, L/2] \text{ for } x^{\text{ld}}_{\text{fail-s}};\ \text{or} && (16)\\
V_{\text{out}} &= V^{\text{out}}_{\text{TH}},\ x_{\text{BD}} \in [L/2, L] \text{ for } x^{\text{ld}}_{\text{fail-d}}. && (17)
\end{aligned}$$

Here, $V^{\text{dr/out}}_{\text{TH}}$ stands for the corresponding threshold voltage (V_H or V_L) of $V_{\text{dr/out}}$.

Algorithm 2 lists the entire procedure for cell-level FP calculation including all four components. The cell-level case index (m, n, k) is determined for NMOS transistor i by finding out the driver cell m, the load cell n and the input pin k.

Algorithm 2 The calculation of cell-level FP using driver and load LUTs. Equation solving uses piecewise-linear approximation based on the LUT data. Failure criteria $V^{\text{dr/out}}_{\text{TH}} = V_H$ or V_L, depends on the nominal values of V_{dr} and V_{out}.

```
1: for each NMOS transistor $i$ in the circuit do
2:     Determine the case index $(m, n, k)$ from $i$
3:     for each input vector $\mathbf{V}$ of this case do
4:         Determine input vectors for driver and load cells: $\mathbf{V}_{\text{dr}}, \mathbf{V}_{\text{ld}}$
5:         {Driver cell $m$ output failure:}
6:         For $x_{\text{BD}} \in [0, L/2]$, obtain $x^{\text{dr}}_{\text{fail-s}}$ as follows
           (if failed, set $x^{\text{dr}}_{\text{fail-s}} = 0$):
           a. Get $I_{\text{TH}} = I_{\text{dr}}(V^{\text{dr}}_{\text{TH}})$ using driver LUT;
           b. Get $x_{\text{BD}}$ by reverse lookup $I_{\text{in}}(V^{\text{dr}}_{\text{TH}}, x_{\text{BD}}) = I_{\text{TH}}$
           using load LUT.
7:         Repeat 6 with $x_{\text{BD}} \in [L/2, L]$, obtain $x^{\text{dr}}_{\text{fail-d}}$
           (If failed, set $x^{\text{dr}}_{\text{fail-d}} = L$).
8:         {Load cell $n$ output failure:}
9:         For $x_{\text{BD}} \in [0, L/2]$, obtain $x^{\text{ld}}_{\text{fail-s}}$ as follows
           (If failed, set $x^{\text{ld}}_{\text{fail-s}} = 0$):
           a. Get subset of $x_{\text{BD}}$ samples, $\mathbf{X}$, satisfying
           $V_{\text{out}}(V^{\text{dr}}_{\text{TH}}, x_{\text{BD}}) \leq V^{\text{out}}_{\text{TH}}$ and $V_{\text{out}}(V^{\text{dr}}_{\text{nom}}, x_{\text{BD}}) \geq V^{\text{out}}_{\text{TH}}$;
           b. For each $x_{\text{BD}} \in \mathbf{X}$, solve (13) for $V_{\text{out}}$, obtain new
           LUT $V_{\text{out}}(x_{\text{BD}})$;
           c. Solve $x_{\text{BD}}$ by reverse lookup $V_{\text{out}}(x_{\text{BD}}) = V^{\text{out}}_{\text{TH}}$
           using the new LUT.
10:        Repeat 9 with $x_{\text{BD}} \in [L/2, L]$, obtain $x^{\text{ld}}_{\text{fail-d}}$
           (If failed, set $x^{\text{ld}}_{\text{fail-d}} = L$).
11:    end for
12:    Calculate $\Pr^{(i)}_{(\text{fail}|\text{BD})}$ using (8), (9) and (10).
13: end for
```

Since the number of V_{in} samples and x_{BD} samples is well bounded, the complexity of solving individual cases is bounded and can be considered as $O(1)$. The calculation of the entire circuit has a linear complexity to the circuit size. In practice, the cost of this is not large, as shown in our simulation results.

In summary, as compared to the direct calculation of cell-level FP which has quadratic complexity as given in (11), the proposed two-stage characterization and cell-level FP scheme effectively reduces both time and space complexity to linear in the library characterization stage, while introducing only a linear-complexity overhead to the circuit analysis stage. This scheme helps keep our entire analysis framework scalable for circuits as well as cell libraries.

IV. Circuit-Level Failure Analysis

Oxide-breakdown-induced logic failure is a weakest-link problem, because failure of any individual logic cell causes the failure of the entire circuit[5]. As shown earlier, prior approaches considered both HBDs and SBDs, and did not adequately differentiate between breakdown events that cause failure and those that do not: in fact, SBD events do not cause functional failures in digital logic circuits [7]. As shown in Section III, some, but not all, HBDs result in circuit failure. Our approach is predicated on identifying the probabilities of HBDs that can cause the circuit to become nonfunctional, and using this information to find the probability of circuit failure with time.

Our novel result on circuit-level FP analysis is stated below, and derives the probability density function of circuit FP based on the parameters of the transistor FP. Specifically, our new result shows that the probability distribution of the time-to-failure for an *entire circuit* is a Weibull distribution. Further, we will see that this implies that the conventional area-scaling based method for circuit FP estimation provides only a loose bound on the time-to-failure. The proof of the result is detailed in Appendix A.

Theorem 1 *The probability distribution $W(t)$, of the time-to-failure, t, for a logic circuit is given by the following distribution:*

$$W(t) = \beta \ln\left(\frac{t}{\alpha}\right) + \ln \sum_{i \in \text{NMOS}} \text{Pr}^{(i)}_{(\text{fail}|\text{BD})} \gamma_i^{\beta} a_i. \tag{18}$$

where α and β are the Weibull parameters for an unit-size device, and $\text{Pr}^{(i)}_{(\text{fail}|\text{BD})}$, γ_i, and a_i are as previously defined in the paper.

This result leads to two important observations.

Observation 1: The time-to-breakdown PDF for a *circuit*, given by (18) is a Weibull distribution. Moreover:

- This distribution has the same Weibull slope, β, as the individual unit-sized device.
- The circuit-level distribution is shifted from that for a unit-sized device. The circuit FP curve is therefore parallel to the transistor FP curve, but is shifted vertically upwards by the *Weibull shift*, defined as:

$$W_{\text{shift}} = \ln \sum_{i \in \text{NMOS}} \text{Pr}^{(i)}_{(\text{fail}|\text{BD})} \gamma_i^{\beta} a_i. \tag{19}$$

  Alternatively, the shift along the horizontal axis shows the logarithm of the lifetime shifted to the left by an amount $\left(-\frac{1}{\beta} \ln \sum \text{Pr}^{(i)}_{(\text{fail}|\text{BD})} \gamma_i^{\beta} a_i\right)$.
- The magnitude of this shift is determined by areas, stress coefficients and cell-level FP of transistors in the circuit.

Observation 2: Our method is more realistic than, and less pessimistic than, the traditional area-scaling-based method for predicting the FP distribution. Specifically, the area-scaling method yields the following Weibull distribution: [1]:

$$W' = \beta \ln\left(\frac{t'}{\alpha}\right) + \ln \sum_{i \in \text{NMOS}} a_i. \tag{20}$$

From (18) and (20), we can obtain that for the same circuit failure $W = W'$,

$$\frac{t}{t'} = \left(\frac{\sum a_i}{\sum \text{Pr}^{(i)}_{(\text{fail}|\text{BD})} \gamma_i^{\beta} a_i}\right)^{\frac{1}{\beta}}. \tag{21}$$

This means our new method shows a relaxation of the circuit lifetime prediction against the traditional area-scaling by a multiplicative factor as given in (21). Since $\text{Pr}^{(i)}_{(\text{fail}|\text{BD})}$ and γ_i are smaller than one, our new method always yields a longer lifetime prediction than the area-scaling approach.

Observation 2 can be interpreted as follows. Unlike the area-scaling-based traditional formula, our result can be considered to use a weighted sum of all areas, or the *effective area*, with the weighting term being $\text{Pr}^{(i)}_{(\text{fail}|\text{BD})} \gamma_i^{\beta}$ for transistor i. This result complies with the intuition that (a) breakdown is slowed by a factor of γ_i, which is equivalent to the area shrinking by γ_i^{β}, (b) for each transistor only breakdowns in certain regions (near source or drain) lead to failure, so the effective area is further decreased by $\text{Pr}^{(i)}_{(\text{fail}|\text{BD})}$ which is actually the worst-case proportion of the failure region.

V. Variation-Aware Oxide Reliability Analysis

While the analysis for the nominal case provides a clear framework for computing the FP, we find (as shown by our results in Section VI) that the effects of variation on the FP are significant. Therefore, in this section, we extend the proposed circuit failure analysis approach to include process variations and spatial correlation. First, we introduce the model for process variations. Next, the transistor-level model and cell-level analysis are updated to capture the effects of variation, and finally, the distribution of circuit failure probability under process variations is derived.

A. Modeling Process Variations

It is widely accepted that process parameter variations can be classified as lot-to-lot, die-to-die (D2D), and within-die (WID) variations, according to their scope; they can also be categorized as systematic and random variations by their causes and predictability. WID variations exhibit spatial dependence knows as spatial correlation, which must be considered for accurate circuit analysis.

[5] Some such failures may lie on false paths and be masked out, but we make the reasonable assumption that the probability that a cell lies on a false path is low, and this scenario can be neglected.

We employ a widely-used variational model: a process parameter X is modeled as a random variable about its mean, X_0, as

$$\begin{aligned} X &= X_0 + X_g + X_s + X_r \\ \sigma_X^2 &= \sigma_{X_g}^2 + \sigma_{X_s}^2 + \sigma_{X_r}^2 \end{aligned} \tag{22}$$

Here, X_g, X_s, and X_r stand for the global part (from lot-to-lot or D2D variations), the spatially correlated part (from WID variation), and the residual random part, respectively. Under this model, all devices on the same die have the same global part X_g. The spatially correlated part is modeled using a method similar as [24], where the entire chip is divided into grids. All devices within the same grid have the same spatially correlated part X_s, and devices in different grids are correlated, with the correlation falling off with the distance. The random part X_r is unique to each device in the system.

In this paper we consider the variations in the transistor width (W), the channel length (L), and the oxide thickness (T_{ox}), and assume Gaussian-distributed parameters. The spatial correlation can be extracted as a correlation matrix [25], and processed using principal components analysis (PCA). The process parameter value in each grid is expressed as a linear combination of the independent principal components, with potentially reduced dimension. For a circuit with n transistors, with the three global parts for W, L and T_{ox}, the spatially correlated part and the n random parts, all the process parameters and their linear functions can be expressed in the random space with basis $\mathbf{e} = [\mathbf{e}_g, \mathbf{e}_s, \epsilon]^{\mathbf{T}}$ as

$$\begin{aligned} X &= X_0 + \Delta X = X_0 + \mathbf{k}_X^{\mathbf{T}}\mathbf{e} \\ &= X_0 + \mathbf{k}_{Xg}^{\mathbf{T}}\mathbf{e}_g + \mathbf{k}_{Xs}^{\mathbf{T}}\mathbf{e}_s + k_\epsilon \epsilon \\ \sigma_X^2 &= \mathbf{k}_X^{\mathbf{T}}\mathbf{k}_X, \;\; \mathrm{cov}(X_i, X_j) = \mathbf{k}_{Xi}^{\mathbf{T}}\mathbf{k}_{Xj} - k_{\epsilon_i}k_{\epsilon_j} \end{aligned} \tag{23}$$

Here, $\mathbf{e}_g = [e_{Wg}, e_{Lg}, e_{Tg}]^{\mathbf{T}}$ is the basis for global part, $\mathbf{e}_s = [e_1, ..., e_t]^{\mathbf{T}}$ is the basis of principal components for the spatial part, and $\epsilon \sim N(0,1)$ is the independent random part for each parameter.

B. Transistor-Level Models under Variations

For transistors with process variations, the Weibull slope β of the time-to-breakdown distribution is a linear function of oxide thickness [4], [26]:

$$\beta_i = \beta_0 + c\ \Delta T_{ox}^{(i)} = \beta_0 + c\ \mathbf{k}_{Ti}^{\mathbf{T}}\mathbf{e} \tag{24}$$

where β_i stands for the Weibull slope for transistor i and β_0 denotes the nominal value. The T_{BD} distribution of i^{th} NMOS transistor under process variation has the same form as (3), with β replaced by β_i. Its area, $a_i = W_i L_i$, is a product of two correlated Gaussians.

The post-breakdown behavior model is also updated to capture the natural randomness of the breakdown resistance, as indicated in Fig. 3(a). The variational models of breakdown resistors, R_s and R_d, are modified to include the variations as follows,

$$\begin{aligned} R_s(x) = R_d(L-x) &= \begin{cases} ae^{bx}(1+\lambda_r\epsilon_r), & 0 \le x \le L_{\text{ext}} \\ kx(1+\lambda_r\epsilon_r), & L_{\text{ext}} \le x \le L \end{cases} \\ \epsilon_r &\sim N(0,1) \end{aligned}$$

This model is consistent with the variations shown in [9].

C. Cell-Level Analysis under Variations

Under process variations, the cell-level failure probability due to a NMOS HBD (taking the breakdown case in Fig. 5 for example) depends on the breakdown resistor and parameters of all transistors in involved driver cell m and load cell n. This dependence is modeled as a linear function of related parameters, using first-order Taylor Expansion. Thus the FP components defined in (9) are updated as

$$\begin{aligned} p &= p_0 + d_r^0\lambda_r\epsilon_r + \sum_j d_{W_j}^0 \Delta W_j \\ &\quad + \sum_j d_{L_j}^0 \Delta L_j + \sum_j d_{T_j}^0 \Delta T_j, \\ p &\in \{p_s^{\text{dr}}, p_s^{\text{ld}}, p_d^{\text{dr}}, p_d^{\text{ld}}\} \end{aligned} \tag{25}$$

Here, d_x^0 is the first-order Taylor coefficients on parameter x. These coefficients are obtained using sensitivity analysis for the cell-level FP characterization, and ΔW_j, ΔL_j and ΔT_j are random variables that can be expressed in the form in (23). Since the FP component p is a linear combination of these process parameters and ϵ_r, it can also be expressed with vector $\mathbf{e}$,

$$\begin{aligned} p &= p_0 + \mathbf{k}_p^{\mathbf{T}}\mathbf{e} + d_r^0\lambda_r\epsilon_r, \\ p &\in \{p_s^{\text{dr}}, p_s^{\text{ld}}, p_d^{\text{dr}}, p_d^{\text{ld}}\} \end{aligned} \tag{26}$$

Note that ϵ_r is the Gaussian representing the breakdown resistor randomness, and is independent of the elements in $\mathbf{e}$.

Using (8), (10), and (26) we can obtain the source-side and drain-side failure probabilities using analytical methods. This involves applying the max operation on correlated Gaussian variables. The work in [27] provided a solution for this max function and approximated the result as a Gaussian in the same random space $\mathbf{e}$. Using such an approach, the final failure probability for case (m,n,k) is calculated by (10) as the sum of two Gaussian variables, and has the form of

$$\text{Pr}_{(\text{fail}|\text{BD})}^{(i)} = \text{Pr}_{(\text{fail}|\text{BD})}^{(m,n,k)} = \text{Pr}_0^{(i)} + \mathbf{k}_{\text{Pr}^{(i)}}^{\mathbf{T}}\mathbf{e} + d_i\epsilon_{r_i} \tag{27}$$

The details of the calculation of failure sensitivities d_x^0's in (25) are given in Appendix B. The characterization and calculation process still maintains linear complexity to the size of library and circuit.

D. Circuit-Level Analysis under Variations

Based on the nominal analysis result (18) of circuit failure probability, we can derive the following under a statistical model:

$$\exp(W) = \sum_{i \in \text{NMOS}} \left(\frac{\gamma_i t}{\alpha}\right)^{\beta_i} \text{Pr}_{(\text{fail}|\text{BD})}^{(i)} a_i \tag{28}$$

Note that $(\frac{t}{\alpha})^{\beta_i}$ is no longer a common factor of the RHS expression due to the device-dependent β_i. Next we define y_i

for each NMOS device i as following

$$\exp(W) = \sum_i \exp(y_i) \tag{29}$$

$$\begin{aligned}\text{where } y_i &= \beta_i \ln\left(\frac{\gamma_i t}{\alpha}\right) + \ln\left(\Pr^{(i)}_{(\text{fail}|\text{BD})} a_i\right) \quad (30)\\ &= \beta_i \ln\left(\frac{\gamma_i t}{\alpha}\right) + \ln \Pr^{(i)}_{(\text{fail}|\text{BD})} + \ln W_i + \ln L_i\end{aligned}$$

Under process variations, for the i^{th} NMOS transistor, β_i is a Gaussian in random space $\mathbf{e}$ as shown in (24); $\Pr^{(i)}_{(\text{fail}|\text{BD})}$ is a Gaussian in space $\mathbf{e} \cup \epsilon_{r_i}$ as in (27); W_i and L_i are also Gaussians in space $\mathbf{e}$ as assumed in Section V-A.

We use two approximations to compute the FP. First, the above logarithms are approximated Gaussians using moment-matching (see Appendix C). As shown in our simulation results section, that approximation does not hurt the final result. Since $\Pr^{(i)}_{(\text{fail}|\text{BD})}$ contains an additional random basis ϵ_{r_i} for breakdown resistor variation, the sum of the logarithms S_i will contain both $\mathbf{e}$ and ϵ_{r_i}. Denoting $\mathbf{k}_{S_i}$ and q_i as the coefficients for these two parts, and μ_{S_i} as the mean of the sum,

$$S_i = \ln \Pr^{(i)}_{(\text{fail}|\text{BD})} + \ln W_i + \ln L_i = \mu_{S_i} + \mathbf{k}^{\mathbf{T}}_{S_i}\mathbf{e} + q_i \epsilon_{r_i} \tag{31}$$

Therefore y_i can be expressed as a Gaussian using $\mathbf{e}$ and ϵ_{r_i}. Denoting $F_i = \ln(\gamma_i t/\alpha)$ and substituting (30) with (31),

$$\begin{aligned} y_i &= \beta_i \ln\left(\frac{\gamma_i t}{\alpha}\right) + S_i \\ &= \beta_{i0} F_i + \mu_{S_i} + (cF_i \mathbf{k}_{T_i} + \mathbf{k}_{S_i})^{\mathbf{T}}\mathbf{e} + q_i \epsilon_{r_i} \quad (32)\end{aligned}$$

which means that y_i is also a Gaussian expressed in terms of $\mathbf{e}$ and ϵ_{r_i}, and $\exp(y_i)$ will have a lognormal distribution. Note that y_i is the Weibull-scale failure probability corresponding to the HBD of i^{th} NMOS transistor.

From (29), $\exp(W)$ is the sum of correlated lognormal RVs. In the second approximation, we model this sum as a lognormal using Wilkinson's method [28], and its first two moments, u_1 and u_2, are[6]

$$\begin{aligned} u_1 &= \sum_i \exp\left(\mu_{y_i} + \sigma^2_{y_i}/2\right) \quad (33)\\ u_2 &= \sum_i \exp\left(2\mu_{y_i} + 2\sigma^2_{y_i}\right) + \\ & \quad 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} e^{\mu_{y_i}+\mu_{y_j}} e^{\frac{1}{2}(\sigma^2_{y_i}+\sigma^2_{y_j}+2r_{ij}\sigma_{y_i}\sigma_{y_j})}\end{aligned}$$

When $\exp(W)$ is small enough, using a first-order Taylor expansion, we find from (40) that

$$\begin{aligned}\Pr^{(\text{ckt})}_{\text{fail}} &= 1 - \exp\left(-\exp(W)\right) \quad (34)\\ &\approx 1 - (1 - \exp(W)) = \exp(W). \quad (35)\end{aligned}$$

This result indicates that, when the circuit failure probability $\Pr^{(\text{ckt})}_{\text{fail}}$ is small (which is actually the case we are interested in, since a circuit with a very large number of breakdowns is unlikely to be functional), it can be approximated with $\exp(W)$, which has lognormal distribution with the first two moments given in (33). When $\Pr^{(\text{ckt})}_{\text{fail}}$ is large, its distribution is unknown, but the mean and variance still can be calculated using a numerical method based on (34). With this information of the circuit failure distribution, it is possible to predict the circuit failure probability at given time t with any specific confidence (e.g. 99%) using the distribution function.

The result also shows that the circuit-level mean-time-to-failure under process variation is no longer a strict Weibull distribution, since the $\sigma^2_{y_i}$ in (33) brings second order term $\ln^2 t$. Although this observation is based on approximations, it is confirmed by simulation results.

Due to the process variations, the mean value of circuit failure probability is increased by the $\sigma^2_{y_i}$ terms in (33). The variance $(u_2 - u_1^2)$ also increases with larger $\sigma^2_{y_i}$. This verifies that process variations exaggerate the likelihood of circuit failure. Moreover, u_2 contains the term r_{ij} which depends positively on the spatial correlation. This means higher spatial correlation will increase the variance of failure probability, thus elevating the reliability issue.

The calculation of y_i in (32) has $O(1)$ complexity due to the limited number of involved devices and principal components. Using the recursive technique proposed in [29], the sum operation over N lognormal variables in (29) can be computed as $N-1$ sum operations on two lognormal variables, keeping the computational complexity at $O(N)$.

VI. Experimental Results

The proposed methods for circuit oxide reliability analysis were applied to the ISCAS85 and ITC99 benchmark circuits for testing. The circuits were synthesized by ABC [30] using the Nangate 45nm open cell library [31], and then placement was carried out using a simulated annealing algorithm. The cell-level library characterization was performed using HSPICE simulation and 45nm PTM model [23]. The circuit-level analysis was implemented in C++ and tested on a Linux PC with 3GHz CPU and 2GB RAM. The parameters for unit-size device the Weibull distribution are $\alpha = 10000$ (arbitrary unit) and $\beta = 1.2$ [4].

A. Results for Nominal Failure Analysis

Three methods for calculating the circuit FP are implemented using a C++ program: (a) Method 1 (M1) performing device-by-device calculation (Equation (36)); (b) Method 2 (M2) using our closed-form formula (Equation (39)); and (c) Monte Carlo (MC) simulation. The implementations of M1 and M2 assume signal independence when computing the stress coefficients, while this is factored into the MC simulation. The MC simulation, performed for each of the time samples, consists of two parts: one, in which the jpmf (see Section II-A) for each transistor stressed in mode A is computed, using 10000 randomized input vectors, and a second, where the breakdown transistors and x_{BD} are randomly generated for 5000 sample circuits, and the probability of circuit failure is computed. For computational efficiency, a biased Monte Carlo

[6] The calculation of u_2 requires the covariance of y_i and y_j. When the HBD case for NMOS i also involves NMOS j (i.e., j belongs to cell m or n) or vice versa, the random parts ϵ of y_i and y_j are actually correlated since they contain process parameters from the same transistor(s). This kind of case is fairly rare (about $2/N$ for a circuit with N logic cells), hence the correlations of the random parts are omitted to simplify the computation.

TABLE I
RUNTIME AND ERROR COMPARISON FOR DIFFERENT METHODS AND DIFFERENT BENCHMARKS, AS WELL AS THE LIFETIME RELAXATIONS.

Circuit	Size	Monte Carlo (MC) Runtime		Method 1 (M1)		Method 2 (M2)		Lifetime
Name	(#Cells)	jpmf	Breakdown	Runtime	$\text{Err}_{\text{M1-MC}}$	Runtime	$\text{Err}_{\text{M2-M1}}$	Relaxation
c432	221	0.39s	9.11s	0.21s	2.37%	10ms	7.51e-5	5.48×
c880	384	0.74s	18.7s	0.34s	2.30%	10ms	2.87e-5	5.50×
c1355	596	1.02s	31.3s	0.29s	2.22%	10ms	2.62e-5	5.34×
c2670	759	1.41s	36.2s	0.83s	3.08%	30ms	2.70e-5	6.16×
c3540	1033	2.55s	67.2s	1.43s	2.21%	60ms	1.27e-5	5.58×
c5315	1699	3.45s	93.9s	1.17s	1.37%	40ms	8.93e-6	5.48×
c6288	3560	17.6s	398s	3.52s	1.74%	130ms	2.93e-6	5.40×
c7552	2316	6.12s	127s	1.69s	1.49%	60ms	5.07e-6	5.29×
b14	4996	35.5s	985s	6.40s	2.81%	250ms	2.09e-6	5.30×
b15	6548	53.3s	2251s	8.53s	1.93%	340ms	2.31e-6	4.83×
b17	20407	209s	8011s	26.6s	3.01%	1060ms	4.56e-7	4.83×
b20	11033	106s	3218s	13.3s	2.06%	530ms	8.01e-7	5.09×
b21	10873	103s	3126s	12.4s	1.69%	490ms	7.78e-7	5.01×
b22	14974	148s	4968s	16.3s	1.16%	650ms	6.34e-7	4.99×

technique is utilized to help the verification for very low circuit FP situations.

Table I presents the detailed runtime and error comparisons for these methods and benchmarks, and shows the lifetime prediction of our method against that of the area-scaling method, as determined by (21). Here, $\text{Err}_{\text{M1-MC}}$ is the error between methods M1 and MC, and $\text{Err}_{\text{M2-M1}}$ is the error between methods M2 and M1. Both errors are measured as the average relative error of FP over a number of time samples. The comparison of M1 with MC shows the effectiveness of the proposed method and demonstrates that the signal independence assumption is appropriate for our benchmarks. The comparison between M2 and M1 validates the approximations made in the proof of Theorem 1. Runtime comparisons (circuit read-in time is not counted in) indicate that the proposed method reduces the runtime by 3 to 4 orders of magnitude, compared with MC. In summary, our new method M2 for circuit failure analysis in (39) is fast and accurate, and it gives a 4.8–6.2× relaxation in the predicted circuit lifetime, as against the traditional area-scaling method.

Fig. 8 visualizes the FP curves for benchmark c7552, which has 2316 cells, as well as the curves using traditional area-scaling and the curve for a unit-size device. The three methods, M1, M2 and MC yield very close results, and all degradation curves share the same Weibull slope. We show a significant relaxation in the circuit lifetime against traditional area-scaling.

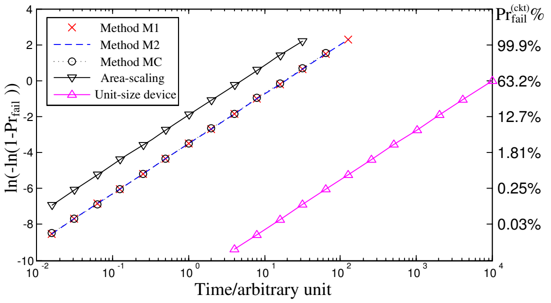

Fig. 8. Result of benchmark circuit c7552 and comparison with traditional area-scaling method and unit-size device.

B. Results for Variation-Aware Failure Analysis

The process variation of T_{ox} is chosen so that its 3σ point is 4% of its mean [14], and is split into 20% of global variation, 20% of spatially correlated variation and 60% of random variation. The variation of W and L sets the 3σ point to 12% of the mean [32], and is split to 40% of global variation, 40% of spatially correlated variation and 20% of random variation. The correlation matrix uses the distance based method in [25]. The number of grids grows with the circuit size.

For each benchmark circuit, the mean and standard deviation of the failure probability are calculated at the time when the nominal circuit has a failure probability of 1%, using the proposed method and Monte Carlo (MC) simulation, separately. The MC simulation randomly generates 5000 circuit instances with different process parameters according to their distribution and correlation models: for each sample, we evaluate the failure probability by using the random value of the process parameters, and performing the nominal analysis described in Sections II to IV.

Table II presents the statistics of the circuit failure probability using the proposed method. The first three columns represent the circuit name and its characteristics. Information about the mean and standard deviation of the failure probability using our approach are presented in the next two columns, and the corresponding relative errors to MC in the following two. It can be seen that our approach closely matches MC, with average errors of 0.72% for the mean and 1.23% for the standard deviation. The value of the mean is very close to the nominal failure probability of 1%, but the standard deviation is considerable. The last two columns compare the circuit lifetime at FP=1% for our approach (using μ+3σ FP) with the nominal approach (using nominal FP) and the area-scaling method under variations (using μ+3σ FP), respectively. We see that the circuit lifetime decrease 19–23% due to process variation, and the proposed approach shows 4.7–5.9× lifetime relaxation against the pessimistic area-scaling method.

Fig. 9 plots the probability density function (PDF) and cumulative density function (CDF) of benchmark c7552 at the nominal failure probability of 1%. The dotted curves show results of MC simulation, while the solid curves show lognormal distribution obtained using proposed method. The

TABLE II
COMPARISONS OF THE MEAN μ AND σ OF CIRCUIT FAILURE.

circuit	Size		Failure probability		Error to MC		Runtime		3σ lifetime	
name	#Cells	#Grids	μ	$\frac{\sigma}{\mu}$	μ	σ	Proposed	MC	Nominal	Area scaling
c432	221	4	1.018%	8.73%	0.18%	0.93%	1.32s	130s	-18.9%	5.23×
c880	384	9	1.024%	8.82%	0.87%	1.52%	1.88s	203s	-19.5%	5.26×
c1355	596	9	1.022%	8.97%	0.11%	0.69%	2.54s	207s	-19.6%	5.16×
c2670	759	16	1.023%	9.10%	0.64%	0.91%	5.70s	532s	-19.9%	5.94×
c3540	1033	16	1.023%	9.34%	0.41%	1.99%	8.16s	842s	-20.3%	5.36×
c5315	1699	25	1.025%	9.49%	0.79%	0.73%	7.75s	743s	-20.6%	5.25×
c6288	3560	64	1.028%	10.4%	0.81%	0.36%	22.7s	2210s	-22.2%	5.23×
c7552	2316	36	1.026%	9.75%	0.73%	0.88%	11.1s	1075s	-21.1%	5.07×
b14	4996	81	1.028%	10.1%	0.56%	1.14%	36.5s	3875s	-21.8%	5.16×
b15	6548	100	1.027%	10.2%	0.52%	0.61%	53.5s	5285s	-21.9%	4.76×
b17	20407	361	1.033%	11.3%	1.13%	0.76%	181s	16634s	-23.8%	4.71×
b20	11033	169	1.031%	10.7%	1.01%	2.75%	80.3s	8100s	-22.8%	4.93×
b21	10873	169	1.031%	10.6%	0.95%	1.32%	73.9s	7593s	-22.7%	4.87×
b22	14974	225	1.032%	10.9%	1.40%	2.65%	104s	10290s	-23.2%	4.84×

nearly perfect match of these two methods validates the approximations made during the analysis, and demonstrates that the circuit failure probability has a lognormal distribution in the region of interest.

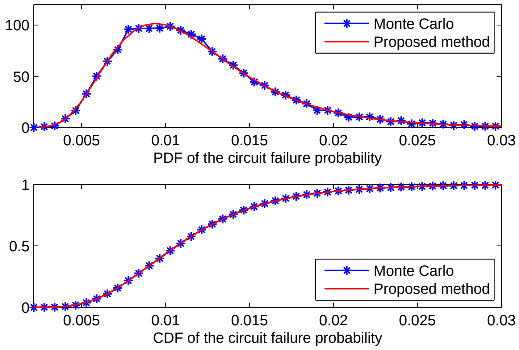

Fig. 9. Comparison of the PDF and CDF of circuit failure.

The proposed method is also tested with other process parameter variance and correlation data besides the condition assumed above. Table III shows the μ+3σ value of circuit failure when nominal circuit failure probability is 1%, and its relative error against MC simulation for benchmark c7552, under several process variation and spatial correlation conditions:

TABLE III
CIRCUIT FAILURE OF C7552 UNDER DIFFERENT TEST CONDITIONS.

Process Variation	Less correlation $g/s/r$=10/10/80%		Medium correlation $g/s/r$=30/40/30%		More correlation $g/s/r$=50/40/10%	
W, L, T_{ox}	μ+3σ	Error	μ+3σ	Error	μ+3σ	Error
σ/μ=1%	1.13%	0.29%	1.23%	0.08%	1.27%	0.09%
σ/μ=2%	1.27%	0.37%	1.47%	0.06%	1.56%	1.60%
σ/μ=5%	1.89%	1.10%	2.48%	0.83%	2.75%	2.58%
σ/μ=10%	4.32%	1.23%	6.57%	3.07%	7.72%	6.23%

The labels g, s, r in the table stand for the global part, the spatially correlated part and the random part of the parameter variations. The results indicate that the relative error to MC simulation is small under all the test conditions, indicating the proposed method is accurate and robust to different conditions of process variations. Moreover, we observe that as the μ+3σ value of the failure probability increases when the process variation increases, or when the correlation increases. This verifies again that the process variations and spatial correlation elevate the reliability issues due to oxide breakdown.

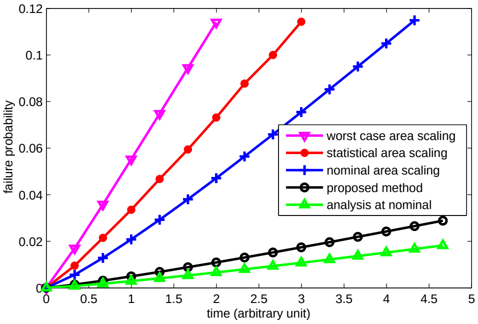

Fig. 10. Comparison of circuit failure, as predicted by various methods.

Finally, Fig. 10 shows a comparison of failure probability vs. time for benchmark c7552 using (a) area scaling with worst-case T_{ox}, (b) area scaling with the T_{ox} variation model in [12], (c) area scaling with nominal T_{ox}, (d) the variation-aware approach proposed in Section V, (e) the analysis method using nominal process parameters as in Section IV. The μ+3σ FP value is used for (b) and (d). The figure leads to several important conclusions. First, it is clear that there are significant differences between area-scaling based methods and our approaches, and that the area-scaling methods are generally too pessimistic. Therefore, to accurately predict circuit reliability, it is essential to account for the inherent circuit resilience and process variations simultaneously. Second, it demonstrates that under either model, the nominal case provides optimistic estimates of the lifetime, and that it is essential to incorporate the effects of variations in order to obtain more accurate lifetime estimates.

VII. CONCLUSION

The paper has focused on the reliability issues caused by gate oxide breakdown in CMOS digital circuits, with the consideration of the inherent resilience in digital circuits that prevents every breakdown from causing circuit failure. The proposed approach takes account for the effective stress for HBD generation and the probability of circuit failure after HBD occurrences. The failure probability for large digital logic circuits is derived in closed form, and it is demonstrated that the circuit-level time-to-failure also follows Weibull distribution and shares the same Weibull slope with the unit-size device. Then the proposed failure analysis approach is extended to include the effect of process variations. The circuit failure probability at specified time instant is derived to be a lognormal distribution due the process variations, and this distribution expands as the process variations and spatial correlation increase. Experimental results show the proposed approaches are effective and accurate compared with Monte Carlo simulation, and give significant better lifetime predictions than the pessimistic area-scaling method.

Appendix A
Proof of Theorem 1

Since failures of different logic cells are independent, the circuit-level FP at time t, $\text{Pr}_{\text{fail}}^{(\text{ckt})}(t)$, can be calculated as:

$$\begin{aligned}\text{Pr}_{\text{fail}}^{(\text{ckt})}(t) &= 1-\prod_{i\in\text{NMOS}}\left(1-\text{Pr}_{\text{fail}}^{(i)}(t)\right)\\ &= 1-\prod_{i\in\text{NMOS}}\left(1-\text{Pr}_{(\text{fail}|\text{BD})}^{(i)}\text{Pr}_{\text{BD}}^{(i)}(t)\right)\end{aligned}$$

Here, $\text{Pr}_{\text{fail}}^{(i)}(t)$ represents the probability that NMOS transistor i in the circuit fails at time t, which implies two facts: first, transistor i breaks down at t, an event that has probability $\text{Pr}_{\text{BD}}^{(i)}(t)$, and second, the breakdown causes a logic failure, which is captured with the cell-level FP $\text{Pr}_{(\text{fail}|\text{BD})}^{(i)}$ from Section III-B. Substituting (3) above:

$$\text{Pr}_{\text{fail}}^{(\text{ckt})}(t) = 1- \prod_{i\in\text{NMOS}}\left(1-\text{Pr}_{(\text{fail}|\text{BD})}^{(i)}\left(1-\exp\left(-\left(\frac{\gamma_i t}{\alpha}\right)^{\beta}a_i\right)\right)\right). \tag{36}$$

This equation gives the circuit FP, incorporating considerations related to the effective stress time and to whether a breakdown event in a transistor causes a cell-level failure. It can further be simplified. For simplicity, we will use the following abbreviated notations: denote $\text{Pr}_{\text{fail}}^{(\text{ckt})}(t)$ by P_f, $\text{Pr}_{(\text{fail}|\text{BD})}^{(i)}$ by p_i, and $(\frac{\gamma_i t}{\alpha})^{\beta}a_i$ by μ_i. Then, taking the logarithm of (36):

$$\ln(1-P_f) = \sum_{i\in\text{NMOS}}\ln\left(1-p_i\left(1-\exp\left(-\mu_i\right)\right)\right). \tag{37}$$

Using first-order Taylor expansions, first using $\exp(-x) = 1-x$ for $x=\mu_i$, and then $\ln(1-x) = -x$ for $x = p_i\mu_i$, the approximation is further simplified as

$$\ln(1-P_f) \approx \sum_{i\in\text{NMOS}}\ln(1-p_i\mu_i) \approx -\sum_{i\in\text{NMOS}}p_i\mu_i. \tag{38}$$

In other words, resubstituting the full forms of P_f, p_i, and μ_i, we get the simplified closed-form formula of the FP as:

$$\text{Pr}_{\text{fail}}^{(\text{ckt})}(t) = 1-\exp\left(-\left(\frac{t}{\alpha}\right)^{\beta}\sum_{i\in\text{NMOS}}\text{Pr}_{(\text{fail}|\text{BD})}^{(i)}\gamma_i^{\beta}a_i\right). \tag{39}$$

For this problem, $0 \le p_i \le 1$ and $0 < \mu_i \ll 1$[7]. Thus the conditions $|x| \le 1, x \ne 1$ for the Taylor expansion of $\ln(1-x)$ are satisfied, and the approximations with first-order Taylor expansions are quite accurate since the high order terms $\text{O}(x^2)$ are much smaller.

We can convert (39) to the following form:

$$\begin{aligned}W &= \ln\left(-\ln\left(1-\text{Pr}_{\text{fail}}^{(\text{ckt})}(t)\right)\right) &(40)\\ &= \beta\ln\left(\frac{t}{\alpha}\right)+\ln\sum_{i\in\text{NMOS}}\text{Pr}_{(\text{fail}|\text{BD})}^{(i)}\gamma_i^{\beta}a_i. &(41)\end{aligned}$$

Appendix B
Cell-Level Characterization under Variations

Under process variations, the I-V characteristics of driver cell and load cell can be expressed using first-order Taylor expansion as

$$\begin{aligned}I_{\text{dr}}(V_{\text{dr}}) &= I_{\text{dr}}^0+\frac{\partial I_{\text{dr}}}{\partial V_{\text{dr}}}\Delta V_{\text{dr}}+\sum_{i\in\text{driver}}\frac{\partial I_{\text{dr}}}{\partial q_i}\Delta q_i &(42)\\ I_{\text{in}}(V_{\text{in}},x_{\text{BD}}) &= I_{\text{in}}^0+\frac{\partial I_{\text{in}}}{\partial V_{\text{in}}}\Delta V_{\text{in}}+\sum_{j\in\text{load}}\frac{\partial I_{\text{in}}}{\partial q_j}\Delta q_j\\ &\quad+\frac{\partial I_{\text{in}}}{\partial \epsilon_r}\epsilon_r+\frac{\partial I_{\text{in}}}{\partial x_{\text{BD}}}\Delta x_{\text{BD}} &(43)\\ V_{\text{out}}(V_{\text{in}},x_{\text{BD}}) &= V_{\text{out}}^0+\frac{\partial V_{\text{out}}}{\partial V_{\text{in}}}\Delta V_{\text{in}}+\sum_{j\in\text{load}}\frac{\partial V_{\text{out}}}{\partial q_j}\Delta q_j\\ &\quad+\frac{\partial V_{\text{out}}}{\partial \epsilon_r}\epsilon_r+\frac{\partial V_{\text{out}}}{\partial x_{\text{BD}}}\Delta x_{\text{BD}} &(44)\end{aligned}$$

Here X^0 denotes the nominal value of parameter X when not considering variations, and q_i, q_j stand for the process parameters (i.e. W, L, and T_{ox}) of the transistors in the driver cell and load cell, respectively. All the first-order derivatives $\partial x/\partial y$ can be calculated in the precharacterization procedure in Algorithm 1 and stored in LUTs.

From (13), we know that $I_{\text{dr}}(V_{\text{dr}}) = I_{\text{in}}(V_{\text{in}},x_{\text{BD}})$, $I_{\text{dr}}^0 = I_{\text{in}}^0$, $V_{\text{dr}} = V_{\text{in}}$, and $\Delta V_{\text{dr}} = \Delta V_{\text{in}}$, therefore from (42) and (43) we get

$$\begin{aligned}&\left(\frac{\partial I_{\text{in}}}{\partial V_{\text{in}}}-\frac{\partial I_{\text{dr}}}{\partial V_{\text{dr}}}\right)\Delta V_{\text{dr}}+\frac{\partial I_{\text{in}}}{\partial x_{\text{BD}}}\Delta x_{\text{BD}}+\frac{\partial I_{\text{in}}}{\partial \epsilon_r}\epsilon_r\\ &\quad+\sum_{j\in\text{load}}\frac{\partial I_{\text{in}}}{\partial q_j}\Delta q_j-\sum_{i\in\text{driver}}\frac{\partial I_{\text{dr}}}{\partial q_i}\Delta q_i = 0 \qquad (45)\end{aligned}$$

To calculate the impact of variations on driver failure, $x_{\text{fail-s}}^{\text{dr}}$ and $x_{\text{fail-d}}^{\text{dr}}$, we have $V_{\text{dr}} = V_{\text{TH}}^{\text{dr}}$, hence $\Delta V_{\text{dr}} = 0$, therefore Δx_{BD} can be solved from (45) as

$$\Delta x_{\text{BD}} = \left(\sum_{i\in\text{driver}}\frac{\partial I_{\text{dr}}}{\partial q_i}\Delta q_i-\sum_{j\in\text{load}}\frac{\partial I_{\text{in}}}{\partial q_j}\Delta q_j-\frac{\partial I_{\text{in}}}{\partial \epsilon_r}\epsilon_r\right)/\frac{\partial I_{\text{in}}}{\partial x_{\text{BD}}}$$

To calculate the impact of variations on load failure, $x_{\text{fail-s}}^{\text{ld}}$ and $x_{\text{fail-d}}^{\text{ld}}$, we have $V_{\text{out}} = V_{\text{TH}}^{\text{out}} = V_{\text{out}}^0$, therefore (44) can be

[7] The region of interest for circuit failure is usually at the lower end, e.g. $P_f < 0.1$. Due to the weakest-link property, the breakdown probability of each individual cell p_i in a large circuit must be very small, which implies that μ_i is very small and must be far less than 1 (considering $\mu_i = 1$ implies that $p_i = 0.632$ for a unit-size device). These approximations are validated by experimental results in Section VI.

rewritten as

$$\frac{\partial V_{\text{out}}}{\partial V_{\text{in}}}\Delta V_{\text{dr}}+\frac{\partial V_{\text{out}}}{\partial x_{\text{BD}}}\Delta x_{\text{BD}}+\sum_{j\in\text{load}}\frac{\partial V_{\text{out}}}{\partial q_j}\Delta q_j+\frac{\partial V_{\text{out}}}{\partial \epsilon_r}\epsilon_r=0 \quad (46)$$

Then using (45) and (46) the unknowns ΔV_{dr} and Δx_{BD} can be solved. The FP components in (25) are obtained using solved Δx_{BD}'s and (9). This variation-aware cell-level analysis approach can be fully integrated to Algorithm 2.

Appendix C
Logarithm of a Gaussian RV

For $x \sim N(\mu_x, \sigma_x^2)$, given $\mu_x \gg \sigma_x > 0$ so that $x > 0$ is always true, its logarithm $y = \ln x$ can be approximated linearly as $y = c + kx$. In order to get better accuracy, the following moment-matching method is used.

For $y = \ln x$, we want to approximate it as $y' {\sim} N(\mu_y, \sigma_y^2)$. Therefore $x' = \exp(y')$ has a lognormal distribution with first two moments

$$\begin{aligned} u_1 &= \exp(\mu_y + \sigma_y^2/2) \\ u_2 &= \exp(2\mu_y + 2\sigma_y^2) \end{aligned} \quad (47)$$

By matching the first two moments of x' and x: $u_1 = \mu_x$, $u_2 = \sigma_x^2 + \mu_x^2$, we can get the distribution of y as

$$\begin{aligned} \mu_y &= 2\ln \mu_x - \frac{1}{2}\ln(\sigma_x^2 + \mu_x^2) \\ \sigma_y^2 &= \ln(\sigma_x^2 + \mu_x^2) - 2\ln \mu_x \end{aligned} \quad (48)$$

Therefore the coefficients for the linear form $y = c + kx$ are $k = \sigma_y/\sigma_x$ and $c = \mu_y - \mu_x\sigma_y/\sigma_x$.

Acknowledgment

The authors would like to thank Professor Montserrat Nafria for initial discussions that led to our exploration of this problem, and the support of the SRC under contract 2007-TJ-1572 and the NSF under award CCR-1017778.

References

[1] J. H. Stathis, "Physical and predictive models of ultrathin oxide reliability in CMOS devices and circuits," *IEEE Trans. Device and Mater. Rel.*, vol. 1, no. 1, pp. 43–59, Mar. 2001.

[2] R. Degraeve, G. Groeseneken, R. Bellens, J. L. Ogier, M. Depas, P. J. Roussel, and H. E. Maes, "New insights in the relation between electron trap generation and the statistical properties of oxide breakdown," *IEEE Trans. Electron Devices*, vol. 45, no. 4, pp. 904–911, Apr. 1998.

[3] J. H. Stathis, "Percolation models for gate oxide breakdown," *J. of Appl. Phys.*, vol. 86, no. 10, pp. 5757–5766, Nov. 1999.

[4] E. Y. Wu, E. J. Nowak, A. Vayshenker, W. L. Lai, and D. L. Harmon, "CMOS scaling beyond the 100-nm node with silicon-dioxide-based gate dielectrics," *IBM J. Res. Dev.*, vol. 46, no. 2/3, pp. 287–298, March/May 2002.

[5] F. Crupi, B. Kaczer, R. Degraeve, A. De Keersgieter, and G. Groeseneken, "A comparative study of the oxide breakdown in short-channel nMOSFETs and pMOSFETs stressed in inversion and in accumulation regimes," *IEEE Trans. Device and Mater. Rel.*, vol. 3, no. 1, pp. 8–13, Mar. 2003.

[6] S. Cheffah, V. Huard, R. Chevallier, and A. Bravaix, "Soft oxide breakdown impact on the functionality of a 40 nm SRAM memory," in *Proc. IRPS*, Apr. 2011, pp. 704–705.

[7] R. Fernández, J. Martin-Martinez, R. Rodriguez, M. Nafria, and X. H. Aymerich, "Gate oxide wear-out and breakdown effects on the performance of analog and digital circuits," *IEEE Trans. Electron Devices*, vol. 55, no. 4, pp. 997–1004, Apr. 2008.

[8] J. Suñé, G. Mura, and E. Miranda, "Are soft breakdown and hard breakdown of ultrathin gate oxides actually different failure mechanisms?" *IEEE Electron Device Lett.*, vol. 21, no. 4, pp. 167–169, Apr. 2000.

[9] R. Degraeve, B. Kaczer, A. De Keersgieter, and G. Groeseneken, "Relation between breakdown mode and location in short-channel nMOSFETs and its impact on reliability specifications," *IEEE Trans. Device and Mater. Rel.*, vol. 1, no. 3, pp. 163–169, Sep. 2001.

[10] S. Tsujikawa, M. Kanno, and N. Nagashima, "Reliable assessment of progressive breakdown in ultrathin mos gate oxides toward accurate TDDB evaluation," *IEEE Trans. Electron Devices*, vol. 58, no. 5, pp. 1468–1475, May 2011.

[11] Y. H. Lee, N. Mielke, M. Agostinelli, S. Gupta, R. Lu, and W. McMahon, "Prediction of logic product failure due to thin-gate oxide breakdown," in *Proc. IRPS*, Mar. 2006, pp. 18–28.

[12] K. Chopra, C. Zhuo, D. Blaauw, and D. Sylvester, "A statistical approach for full-chip gate-oxide reliability analysis," in *Proc. ICCAD*, Nov. 2008, pp. 698–705.

[13] B. Kaczer, R. Degraeve, M. Rasras, K. Van de Mieroop, P. J. Roussel, and G. Groeseneken, "Impact of MOSFET gate oxide breakdown on digital circuit operation and reliability," *IEEE Trans. Electron Devices*, vol. 49, no. 3, pp. 500–506, Mar. 2002.

[14] Z. Cheng, D. Blaauw, and D. Sylvester, "Post-fabrication measurement-driven oxide breakdown reliability prediction and management," in *Proc. ICCAD*, Nov. 2009, pp. 441–448.

[15] F. N. Najm, "A survey of power estimation techniques in VLSI circuits," *IEEE Trans. VLSI Syst.*, vol. 2, no. 4, pp. 446–455, Dec. 1994.

[16] R. Burch, F. N. Najm, P. Yang, and T. N. Trick, "A Monte Carlo approach for power estimation," *IEEE Trans. VLSI Syst.*, vol. 1, no. 1, pp. 63–71, Mar. 1993.

[17] R. Rodríguez, J. H. Stathis, and B. P. Linder, "A model for gate-oxide breakdown in CMOS inverters," *IEEE Electron Device Lett.*, vol. 24, no. 2, pp. 114–116, Feb. 2003.

[18] H. Wang, M. Miranda, F. Catthoor, and D. Wim, "Impact of random soft oxide breakdown on SRAM energy/delay drift," *IEEE Trans. Device and Mater. Rel.*, vol. 7, no. 4, pp. 581–591, Dec. 2007.

[19] B. Kaczer, R. Degraeve, A. De Keersgieter, K. Van de Mieroop, V. Simons, and G. Groeseneken, "Consistent model for short-channel nMOSFET after hard gate oxide breakdown," *IEEE Trans. Electron Devices*, vol. 49, no. 3, pp. 507–513, Mar. 2002.

[20] J. Segura, C. Benito, A. Rubio, and C. F. Hawkins, "A detailed analysis and electrical modeling of gate oxide shorts in MOS transistors," *J. of Electron. Test.*, vol. 8, no. 3, pp. 229–239, Jun. 1996.

[21] X. Lu, Z. Li, W. Qiu, D. M. H. Walker, and W. Shi, "A circuit level fault model for resistive shorts of MOS gate oxide," in *Proc. WFOPC*, Sep. 2004, pp. 97–102.

[22] K. Shubhakar, K. L. Pey, S. S. Kushvaha, M. Bosman, S. J. O'Shea, N. Raghavan, M. Kouda, K. Kakushima, Z. R. Wang, H. Y. Yu, and H. Iwai, "Nanoscale electrical and physical study of polycrystalline high-k dielectrics and proposed reliability enhancement techniques," in *Proc. IRPS*, Apr. 2011, pp. 786–791.

[23] Predictive Technology Model. Available: http://www.eas.asu.edu/~ptm/.

[24] H. Chang and S. S. Sapatnekar, "Statistical timing analysis considering spatial correlations using a single PERT-like traversal," in *Proc. ICCAD*, Nov. 2003, pp. 621–625.

[25] J. Xiong, V. Zolotov, and L. He, "Robust extraction of spatial correlation," *IEEE Trans. Comput.-Aided Des.*, vol. 26, no. 4, pp. 619–631, Apr. 2007.

[26] R. Degraeve, G. Groeseneken, R. Bellens, M. Depas, and H. E. Maes, "A consistent model for the thickness dependence of intrinsic breakdown in ultra-thin oxides," in *Proc. IEDM*, Dec. 1995, pp. 863–866.

[27] C. Clark, "The greatest of a finite set of random variables," *Operations Research*, vol. 9, pp. 85–91, 1961.

[28] A. A. Abu-Dayya and N. C. Beaulieu, "Comparison of methods of computing correlated lognormal sum distributions and outages for digital wireless applications," in *Proc. VTC*, vol. 1, Jun. 1994, pp. 175–179.

[29] A. Srivastava, S. Shah, K. Agarwal, D. Sylvester, D. Blaauw, and S. Director, "Accurate and efficient gate-level parametric yield estimation considering correlated variations in leakage power and performance," in *Proc. DAC*, 2005, pp. 535–540.

[30] Berkeley Logic Synthesis and Verification Group, "Abc: A system for sequential synthesis and verification, release 70930." Available: http://www.eecs.berkeley.edu/~alanmi/abc/.

[31] Nangate 45nm Open Cell Library. Available: http://www.nangate.com/.

[32] International technology roadmap for semiconductors, 2008 update, process integration, devices and structures.