Scalable Methods for the Analysis and Optimization of Gate Oxide Breakdown

Jianxin Fang, Sachin S. Sapatnekar
Department of ECE, University of Minnesota
{fang0116,sachin}@umn.edu

Abstract

In this paper we first develop an analytic closed-form model for the failure probability (FP) of a large digital circuit due to gate oxide breakdown. Our approach accounts for the fact that not every breakdown leads to circuit failure, and shows a 6–11× relaxation of the predicted lifetime with respect to the ultra-pessimistic area-scaling method. Next, we develop a posynomial-based optimization approach to perform gate sizing for oxide reliability in addition to timing and area.

1. Introduction

Oxide breakdown in CMOS circuits 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 SiO_2 gate oxide becomes conductive when a critical defect density is reached at a certain location in the oxide. With device scaling, electric fields across the gate oxide have progressively increased as supply voltages have scaled down slower than the oxide thickness. This makes the transistors more vulnerable to breakdown, and as a result, more susceptible to failures due to oxide breakdown.

At the device level, the mechanism 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, 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 [2,3].

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. 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). SBDs can cause parametric variations but not functional failures [4], and are not considered here. Through the rest of this paper, the term "circuit failure" will mean a functional failure of the circuit.

As documented in the literature, it is believed that there is no substantial difference between the physical origins of these two breakdown modes, 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 [5].

At the circuit level, the traditional failure prediction method for a large circuit uses area-scaling, extrapolated from single device characterization [1, 2]. 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 [6], or by considering the variation of gate-oxide thickness [7]. 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 2 and 3.

Precise analysis or measured results on very small circuits, such as op amps, nor gates, ring oscillators, and dynamic gates, have been published [8], based on the post-breakdown behavior models. Some of these works showed that digital circuits can survive several hard breakdowns without losing the functionality [9]. These methods, either using complex analysis models or based on measurements, cannot easily be extended to general large-scale digital circuits in a computationally scalable manner.

The contribution of our work is twofold. First, we develop a scalable method for analyzing the failure probabilty (FP) of large digital circuits, while realistically considering the circuit environment that leads to stress and oxide breakdown. To the best of our knowledge, currently published work can only successfully perform this analysis on very small circuits, or uses gross approximations for large circuits. To achieve this goal, at the *transistor* level, we revise the Weibull time-to-breakdown model to incorporate the actual stressing 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 [5]. 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. 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.

Second, we use our model to develop an optimization approach to mitigate the effect of gate oxide breakdown. We demonstrate that by appropriately sizing the devices, the circuit can be made more resilient, so that it performs correctly even in the presence of oxide breakdown events. We formulate a problem that performs transistor sizing with the aim of increasing the time to circuit failure, while addressing conventional sizing goals such as power and delay. Experimental results show that circuit reliability can be improved by increasing the area, which runs counter to the prediction of the traditional area-scaling theory.

2. Transistor-Level Models

In this section, we discuss models for the time-to-breakdown and the post-breakdown behavior of a transistor. Sections 2.1 and 2.2 largely overview existing models, while Section 2.3 presents our new simple quantitative model for breakdown resistance that can be calibrated from experimental data. Our primary focus here is on SiO_2-based dielectrics for which published data are available in the public domain. However, the proposed methodology is applicable to circuit-level analysis for circuits using high-K dielectrics, which are also susceptible to dielectric breakdown.

Our discussion is guided by two observations:

- Only hard breakdowns cause serious device degradations [1].
- The occurrence of hard breakdown is very prevalent in NMOS transistors but relatively rare in PMOS devices [3].

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

Furthermore, our work assumes that a transistor will be affected by at most one HBD. This assumption is reasonable: 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[1]. This assumption is similar in spirit to the single stuck-at fault assumption in the test arena.

[1] It can be shown that this probability is around 3.24e-5 when the circuit has FP of 0.1 for benchmark c7552 under our experiment conditions in Section 6.

2.1 Time to Breakdown

The transistor time-to-breakdown, T_{BD}, is typically treated statistically using a Weibull distribution, with an area-scaling formula [2]. The breakdown probability of device i, with area a_i, at time t is

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

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. Plotting $W = \ln(-\ln(1 - \Pr_{BD}^{(i)}(t)))$ against $\ln(t)$ yields a straight line with slope β, and this is commonly referred to as the Weibull plot.

The parameters α and β in (1) are usually characterized in experiments, as described in [2, 5], 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 digital circuits. Typically, in a circuit setting, the logic states at the transistor terminals change over time, with six possible stress modes for a NMOS transistor, as shown in Figure 1[2].

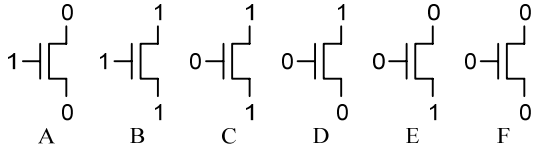

Figure 1: Stressing modes for NMOS transistors.

An HBD occurs in the case of NMOS stressed in inversion, while an NMOS in accumulation almost always experiences SBD [3]. In Figure 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 hard breakdowns in a device, and potential circuit failure. We introduce the stressing coefficient, γ_i, for device i to capture the proportion of this effective stress time, and reformulate Equation (1) as

$$\Pr_{BD}^{(i)}(t) = 1 - \exp\left(-\left(\frac{\gamma_i t}{\alpha}\right)^{\beta} a_i\right) \quad (2)$$

where $(\gamma_i t)$ represents the effective stressing period after time t of circuit operation. The stressing 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), or (g,s,d), 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.

2.2 Post-Breakdown Behavior

Figure 2 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 overlap regions between gate and source/drain are assumed to be of length L_{ext}. The distance from the source is denoted by x, and the breakdown location is assumed to be at x_{BD}.

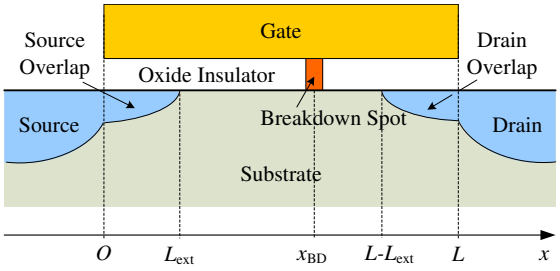

Figure 2: Schematic of oxide breakdown in a transistor.

[2]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.

Various modeling approaches for post-breakdown analysis at the transistor- or cell-level have been proposed in the literature. The work in [10] suggests a complex physical model that reduces to a simple resistor model when the breakdown location is near the source or drain. Independent experiments have reported that HBDs show a roughly linear (ohmic) I-V characteristic [1]. Based on this, we use a simpler linear resistor model, similar to that in [9], 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 Figure 3(a). The values of these two resistors are dependent on the breakdown location, x_{BD}.

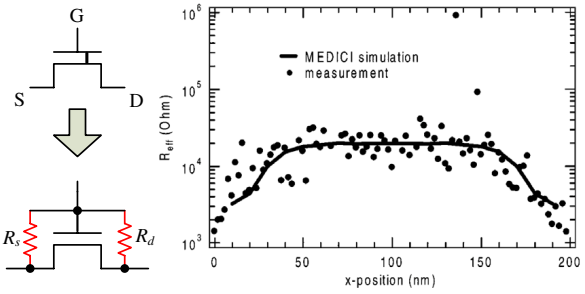

Figure 3: (a) Resistor model of post-breakdown behavior. (b) The effective resistance as a function of breakdown location [5].

In characterizing the values of these resistances, it is important to lay down some requirements that they must fulfill. Figure 3(b) 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 [5]. The data points in this figure correspond to measurements, while the solid line is based on a detailed device simulation. Further experimental data in [5] (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 [5] are that:

- R_{BD} is smaller when the breakdown occurs in the source or drain overlap 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.

2.3 Modeling the Breakdown Resistors

While the structure of the breakdown resistor model using R_s and R_d in Figure 3(a) is not fundamentally new, there has been less work on deriving a model that relates R_{BD} with x_{BD}, since this relationship is very important for statistical modeling. The only known work is an equivalent circuit model in [10], but it requires a complex characterization process; moreover, the nonlinearity of the model makes its evaluation in a circuit simulator more computational. 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 R_{BD} vs. x_{BD} curve in Figure 3(b). We propose to capture the variation of R_{BD} with x_{BD} through a piecewise linear/log-linear model, where R_s [R_d] varies exponentially with x_{BD} in the source [drain] overlap region, and linearly in the remainder of the channel:

$$R_s(x) = \begin{cases} kx, & L_{ext} \le x \le L \\ ae^{bx}, & 0 \le x \le L_{ext} \end{cases} \quad (3)$$

Due to source-drain symmetry[3], we obtain $R_d(x) = R_s(L - x)$. When

[3]For asymmetric transistors, the ideas of this work can easily be extended for a similar modeling and characterization scheme.

both the source and drain nodes are grounded,

$$R_{\mathrm{BD}}(x) = R_s(x) \parallel R_d(x) \tag{4}$$

The value of R_{BD} is at its minimum, $R_{\mathrm{BD\,min}}$, at $x = 0$ and $x = L$, and by symmetry, at its maximum, $R_{\mathrm{BD\,max}}$ at $x = L/2$.

The constants k, a and b are obtained from measurements by matching a set of boundary conditions.

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

Four parameters are required to characterize this model: L, L_{ext}, $R_{\mathrm{BD\,min}}$ and $R_{\mathrm{BD\,max}}$. Figure 4 shows an example plot for R_{BD} using this model, with the parameters L = 45nm, L_{ext} = 13nm, $R_{\mathrm{BD\,max}}$ = 20kΩ, and $R_{\mathrm{BD\,min}}$ = 1kΩ. It is easy to see that the results here are well matched to the trend of experimental results in Figure 3(b).

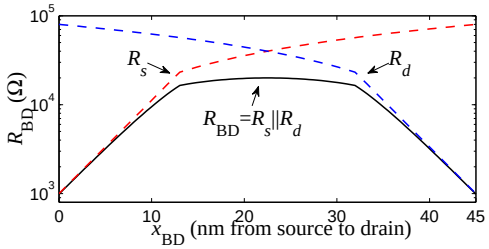

Figure 4: Breakdown resistor as a function of location.

3. Cell-Level Failure Analysis

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 logic circuits.

3.1 Breakdown Case 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). Figure 5(a) shows an example of such a breakdown case[4], using a NAND2 as cell m, a NOR2 as cell n, and k = 1.

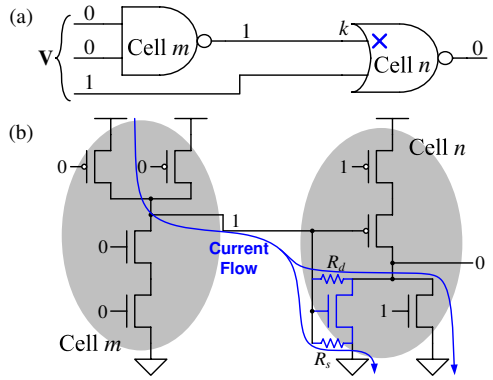

Figure 5: Cell-level analysis of the breakdown case.

[4]The probability that cell m or cell n contains other broken down transistors is quite small (about 2.14e-4 when the circuit has FP of 0.1 for benchmark c7552 under our experiment conditions), thus negligible.

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 Figure 5(a), q = 3, and we consider the assignment $\mathbf{V} = (0, 0, 1)$. We refer to a breakdown case for a 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 2. The transistor-level circuit, using the resistor model, is shown in Figure 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.

3.2 Calculation of Failure Probabilities (FPs)

The breakdown case in Figure 5 is analyzed using SPICE DC sweep under 45nm PTM models [11] and V_{dd} = 1.2V. The steady-state output voltages of cells m and n, as a function of x_{BD}, are shown in Figure 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 are likely to shift away from their nominal values of V_{dd} and 0, respectively. When the voltages go beyond certain limits, the logic could flip and result in circuit failure.

Note that the results for cells m and n are asymmetric for the input excitation in Figure 5, in that m shows a failure when the defect lies at either end of the channel, while the failure for 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, $V_H = 0.7V_{dd}$, $V_L = 0.3V_{dd}$), 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 monotonicaly 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[5]. We define these intervals for gate g to be $[0, x^{(g)}_{\text{fail-d}}]$ and $[x^{(g)}_{\text{fail-s}}, L]$, respectively, at the drain and source end.

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

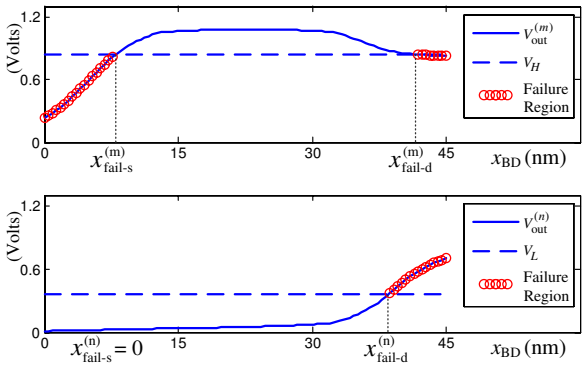

Figure 6: Cell output voltages under breakdown.

We can then obtain the souce-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 [5], since the breakdown position is uniformly distributed in the channel, i.e., $x_{\mathrm{BD}} \sim \mathrm{U}[0, L]$. Therefore, these

[5]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 Figure 6.

FPs are given by:

$$\Pr^{(m,n,k,\mathbf{V})}_{(\text{fail-s}|\text{BD})} = \frac{1}{L}\max\left(x^{(m)}_{\text{fail-s}}, x^{(n)}_{\text{fail-s}}\right) \quad (5)$$

$$\Pr^{(m,n,k,\mathbf{V})}_{(\text{fail-d}|\text{BD})} = \frac{1}{L}\left(L - \min\left(x^{(m)}_{\text{fail-d}}, x^{(n)}_{\text{fail-d}}\right)\right)$$

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:

$$\Pr^{(m,n,k)}_{(\text{fail}|\text{BD})} = \max_{\mathbf{V}\in\mathbb{B}^q} \Pr^{(m,n,k,\mathbf{V})}_{(\text{fail-s}|\text{BD})} + \max_{\mathbf{V}\in\mathbb{B}^q} \Pr^{(m,n,k,\mathbf{V})}_{(\text{fail-d}|\text{BD})} \quad (6)$$

Since the logic cells come from a common cell library, $\mathbb{C}$, it is possible to characterize a library over all breakdown cases (m, n, k) as a precomputation with complexity $O(|\mathbb{C}|^2)$. For circuit-level failure analysis, as described in Section 4, the precomputed FP results can be retrieved from the characterized library in $O(1)$ time. Some special scenarios are easily handled during analysis: multifinger devices can be treated as a single area-scaled larger device; gates with multiple transistors connected to one input pin can be modeled as a single transistor with an equivalent area.

4. 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[6]. Prior approaches did not adequately differentiate between breakdown events that cause failure and those that do not. As shown in Section 3, 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 circuit FP 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 nontrivial, but to maintain the flow of this paper, it is detailed in the Appendix.

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

$$W(t) = \beta \ln\left(\frac{t}{\alpha}\right) + \ln \sum_{i\in\text{NMOS}} \Pr^{(i)}_{(\text{fail}|\text{BD})} \gamma_i^\beta a_i. \quad (7)$$

where α and β are the Weibull parameters for a unit-size device, and $\Pr^{(i)}_{(\text{fail}|\text{BD})}$, 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 Equation (7) 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:

$$W_{\text{shift}} = \ln \sum_{i\in\text{NMOS}} \Pr^{(i)}_{(\text{fail}|\text{BD})} \gamma_i^\beta a_i. \quad (8)$$

  Alternatively, the shift along the horizontal axis shows the logarithm of the lifetime shifted to the left by $\left(-\frac{1}{\beta}\ln\sum \Pr^{(i)}_{(\text{fail}|\text{BD})}\gamma_i^\beta a_i\right)$.

[6] 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 can be neglected.

- The magnitude of this shift is determined by areas, stressing 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 failure probability distribution. Specifically, the area-scaling method yields the following Weibull distribution: [1]:

$$W' = \beta \ln\left(\frac{t'}{\alpha}\right) + \ln \sum_{i\in\text{NMOS}} a_i. \quad (9)$$

From Equations (7) and (9), we can obtain that for the same circuit failure $W = W'$, our new method shows a relaxation of the predicted circuit lifetime against the traditional way by a multiplicative factor of $(\sum a_i / \sum \Pr^{(i)}_{(\text{fail}|\text{BD})}\gamma_i^\beta a_i)^{1/\beta}$.

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 $\Pr^{(i)}_{(\text{fail}|\text{BD})}\gamma_i^\beta$ 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 $\Pr^{(i)}_{(\text{fail}|\text{BD})}$ which is actually the worst-case proportion of the failure region.

5. Gate Sizing for Reduced Failure Probability

The circuit level failure analysis in Section 4 shows that for a circuit designed in a given technology, the FP is affected by the Weibull shift, W_{shift}, given by Equation (8).

We define the *lifetime* of a circuit as the time corresponding to a specified failure probability, W. In other words, this is the time at which the right hand side of Equation (7) evaluates to W. It is easy to show that under this failure probability, if the Weibull shift for a circuit is reduced from $W^{(0)}_{\text{shift}}$ to $W^{(1)}_{\text{shift}}$, then the impact on the circuit lifetime is given by the following exponential relationship:

$$\frac{t_1}{t_0} = \exp\left(\frac{1}{\beta}\left(W^{(0)}_{\text{shift}} - W^{(1)}_{\text{shift}}\right)\right) \quad (10)$$

Therefore by reducing the Weibull shift, it is possible to lower the FP and prolong the lifetime of the circuit.

We achieve this through gate sizing, an optimization that traditionally explores area/delay/power tradeoffs by sizing the logic cells in the circuit [12]. We will next demonstrate how the Weibull shift is affected by the sizes of the logic cells in the circuit, and use it to build a framework for reliability-driven gate sizing.

5.1 Modeling of the Weibull Shift

From Section 3, the cell-level FP, $\Pr^{(i)}_{(\text{fail}|\text{BD})}$, is obtained by analyzing the breakdown case of cells m and n (Figure 5). Therefore it depends on the sizes of these cells and can be represented as:

$$\Pr^{(i)}_{(\text{fail}|\text{BD})} = f(s_m, s_n), \quad (11)$$

where s_m and s_n are the sizing factors for cells m and n, i.e., the multiples of their sizes with respect to their nominal sizes. Clearly, the area of an nmos transistor i, $a_i = s_n a_{i(\text{nominal})}$, and this depends on s_n. Therefore, we define a set of new functions $Q^{(i)}$ to include all the sizing-dependent elements in Equation (8):

$$Q^{(i)}(s_m, s_n) = \Pr^{(i)}_{(\text{fail}|\text{BD})} s_n = s_n f(s_m, s_n). \quad (12)$$

The Weibull shift of the circuit can be rewritten as

$$W_{\text{shift}} = \ln \sum_{i\in\text{NMOS}} Q^{(i)}(s_{m(i)}, s_{n(i)}) \gamma_i^\beta a_{i(\text{nominal})}, \quad (13)$$

where $n(i)$ $[m(i)]$ refers to the logic cell that contains [drives] the i^{th} NMOS transistor.

The computation of the $Q^{(i)}$ functions requires the calculation of FP $\Pr^{(i)}_{(\text{fail}|\text{BD})}$, which does not admit a simple closed form. Therefore, to find the $Q^{(i)}(s_m, s_n)$ function for each breakdown case, we perform SPICE-based analysis as a numerical alternative. For each case $i \to (m, n, k)$, the $Q^{(i)}(s_m, s_n)$ function is computed with a set of sampled s_m and s_n values and stored in a look-up table during library characterization.

5.2 Reliability-Driven Gate Sizing

In order to take circuit failure into consideration, we can add a new constraint for the Weibull shift to the sizing problem, to limit the shift in the Weibull curve, $W_{\text{shift}} \le W_{\max}$, where $W_{\max}$ denotes the maximum acceptable Weibull shift under a circuit lifetime spec.

The conventional gate sizing problem is usually solved using geometric programming (GP) , in which the objective and constraints are modeled using posynomials, and the problem is then transformed to a convex optimization problem and solved by standard solvers. However the Weibull shift function, a weighted sum of Q functions of all transistors, cannot be directly represented as a posynomial of the sizing factors. To address this problem and adapt the Weibull shift constraint into the GP framework, an empirical generalized posynomial fit for the Q functions is proposed:

$$Q_{fit} = \max(Q_{f1}, Q_{f2}) - q, \tag{14}$$

$$\text{where} \quad Q_{f1} = c_1 \left(\frac{s_n}{s_m}\right)^{b_1}; \; Q_{f2} = c_2 \left(\frac{1}{s_m}\right)^{b_2} + d; \; b_1, b_2, c_1, c_2, d, q \ge 0.$$

Here Q_{fit} is the maximum of two posynomial functions. Experimental results show a 5.82% average relative error of fitting for the tested library in Section 6. Since all fitting parameters are non-negative, $Q_{fit} + q$ is a generalized posynomial.

Based on the proposed model, we define intermediate variables $Q_m = \max(Q_{f1}, Q_{f2})$ to ensure the posynomial property, and use Q_{fit} to replace Q in Equation (13) to obtain

$$\exp(W_{\text{shift}}) = \sum_{i \in \text{NMOS}} Q_m^{(i)} \gamma_i^{\beta} a_{i(\text{nominal})} - \sum_{i \in \text{NMOS}} q_i \gamma_i^{\beta} a_{i(\text{nominal})}. \tag{15}$$

The constraint $W_{\text{shift}} \le W_{\max}$ can now be rewritten as

$$\begin{aligned} \sum_{i \in \text{NMOS}} Q_m^{(i)} \gamma_i^{\beta} a_{i(\text{nominal})} &\le \exp(W_{\max}) + \sum_{i \in \text{NMOS}} q_i \gamma_i^{\beta} a_{i(\text{nominal})}, \\ Q_{f1}^{(i)} / Q_m^{(i)} &\le 1, \quad i \in \text{NMOS}, \\ Q_{f2}^{(i)} / Q_m^{(i)} &\le 1, \quad i \in \text{NMOS}. \end{aligned} \tag{16}$$

Note that all right hand sides above are constants, and these constraints are in posynomial form and can directly be applied to the conventional sizing problem. The new problem, containing the Weibull shift constraints, can be solved by traditional GP solvers.

Due to the nonconvex property of the original Weibull shift function, it is difficult to find the global optimum of the sizing problem. The newly proposed posynomial fit for Q functions adjusts the search space to a convex set, with minimal loss in accuracy. Thus the global optimum of the modified problem can be regarded as a close approximation for the solution of the original problem.

6. Experimental Results

Our methods were applied to the ISCAS85 and ITC99 benchmark circuits for testing on a Linux PC with 3GHz CPU. The library characterization is performed with HSPICE using models and parameters described in the previous sections. Parameters for unit-size device Weibull distributions are $\alpha = 10000$ and $\beta = 1.2$.

6.1 Results for Failure Analysis

For failure analysis, the benchmark circuits were synthesized with SIS using a library consisting of 40 logic cells, including inverters of 10 different sizes, and NAND2, NAND3, NOR2, NOR3, AOI3, and OAI3, with 5 different sizes for each kind of cell.

Three methods for calculating the circuit FP are implemented using a C++ program: (a) Method 1 (M1) performing device-by-device calculation (Equation (17)); (b) Method 2 (M2) using our closed-form formula (Equation (21)); and (c) Monte Carlo (MC) simulation. The implementations of M1 and M2 assume signal independence when computing the stressing coefficients, while this is factored into the MC simulation.

Table 1 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. Here, $E_{\text{M1-MC}}$ is the error between methods M1 and MC, and $E_{\text{M2-M1}}$ 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 4 to 5 orders of magnitude, compared with MC. In summary, our new method M2 for circuit failure analysis in Equation (21) is fast and accurate, and it gives a 6–11× relaxation in the predicted circuit lifetime, as against the traditional area-scaling method.

Table 1: Runtime and error comparison, and the lifetime relaxations.

circuit Name	Size (#cell)	MC runtime	Method 1 (M1)		Method 2 (M2)		Life Relax
			runtime	$E_{\text{M1-MC}}$	runtime	$E_{\text{M2-M1}}$	
c432	366	18.6s	10ms	6.00%	<10ms	2.07e-4	11.2×
c880	474	31.1s	10ms	2.88%	<10ms	1.29e-4	6.77×
c2670	1173	75.7s	20ms	2.36%	<10ms	3.95e-5	6.67×
c3540	1521	122s	40ms	1.93%	<10ms	2.50e-5	6.59×
c5315	2479	236s	50ms	2.41%	<10ms	1.64e-5	6.56×
c6288	2696	186s	60ms	1.95%	<10ms	8.07e-6	5.95×
c7552	3960	778s	100ms	3.50%	10ms	1.02e-5	7.56×
b14	10136	3699s	270ms	2.15%	20ms	2.48e-6	10.0×
b15	14843	6228s	360ms	5.52%	30ms	1.07e-6	7.24×
b17	38741	19060s	1160ms	2.93%	170ms	5.20e-7	8.66×
b20	18886	7171s	750ms	2.88%	70ms	1.03e-6	8.60×
b21	15917	5989s	400ms	6.34%	70ms	9.15e-7	7.80×
b22	20595	11271s	540ms	2.08%	80ms	4.92e-7	6.15×

Figure 7 shows FP vs. time for benchmark c7552 using our method as well as traditional area-scaling, and the curve for a unit-size device. M1, M2 and MC yield very close results, all degradation curves share the same Weibull slope, and our method significantly reduces the pessimistic lifetime predictions from traditional area-scaling.

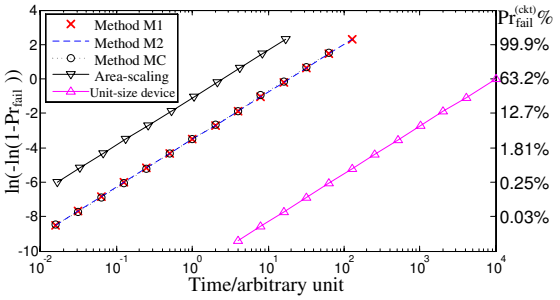

Figure 7: Result of benchmark circuit c7552 and comparison with traditional area-scaling method and unit-size device.

6.2 Results for Gate Sizing

For reliability-driven gate sizing, we work with a library that is characterized by calculating the Q function with sampled sizing factors for all breakdown cases, and then fitting the Q functions using Matlab for each case. For the total of 119 cases, there is a 5.82% average relative error of fitting (RMSE divided by the maximum of Q function, then averaged for all cases).

The benchmark circuits were mapped to this library. Then we use transistor area and delay models consistent with [12] and Mosek [13] Optimization Toolbox for Matlab as the GP solver.

To verify the usefulness of gate sizing for reliability, each of the benchmark circuits is first optimized for delay without a W_{shift} constraint, to obtain the minimum delay d_0 and corresponding area a_0. The corresponding unoptimized value of W_{shift} at this minimum delay point is shown in the second column of Table 2. The circuit is then optimized to minimize W_{shift}, subject to a delay constraint of $1.1\times d_0$ and an area constraint of a_0. The solution, listed in the third column, shows the W_{shift} improvement at the cost of 10% more delay. The fourth column lists the corresponding lifetime improvement calculated using Equa-

tion (10). For the fifth column, the area constraint is loosened to $2a_0$, for further improvement of W_{shift}, and the corresponding lifetime improvement is provided in the last column. Over all tested benchmarks, the results show 1.1–1.5× lifetime improvement when the delay constraint is relaxed to 1.1× of the minimum delay, and another 1.2–1.9× improvement when an additional 2× area is allowed.

Table 2: Lifetime improvement by gate sizing.

Circuit Name	W_{shift} at min delay d_0, a_0 (I)	min W_{shift} $D \le 1.1d_0$ $A \le a_0$ (II)	Lifetime Improve II vs. I	min W_{shift} $D \le 1.1d_0$ $A \le 2a_0$ (III)	Lifetime Improve III vs. II
c432	6.01	5.52	1.50×	5.12	1.40×
c880	5.98	5.83	1.13×	5.27	1.59×
c2670	7.02	6.80	1.20×	6.45	1.33×
c3540	7.47	7.14	1.31×	6.76	1.37×
c5315	7.91	7.66	1.23×	7.40	1.24×
c6288	7.95	7.67	1.26×	6.92	1.87×
c7552	8.23	8.06	1.16×	7.81	1.23×

As a typical gate sizing example, Figure 8 presents the area vs. Weibull shift trade-off curves under different delay constraints for benchmark circuits c880, c2670, and c3540. The triangle points in the plot indicate the area a_0 and W_{shift} at minimum delay for each circuit. Two curves under different delay constraints are plotted for each circuit. The x-axis shows both W_{shift} and the absolute lifetime when circuit FP = 5%. The figure shows that the circuits sized for minimum delay generally have the worst lifetime values, i.e., the triangles are to the right of the curves, and by loosening the delay and/or area constraints, the lifetime can be improved.

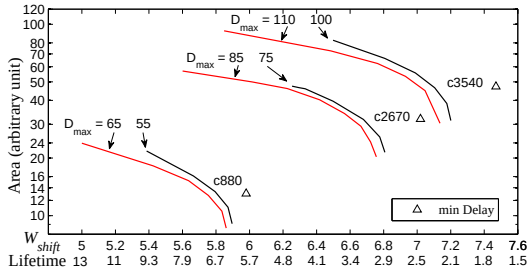

Figure 8: The area vs. Weibull shift trade-off curves.

We have shown that *circuit reliability can be improved by increasing the area, which runs counter to the prediction of the traditional area-scaling theory* of Equation (9), which claims higher FP for larger circuit size. This apparent contradiction can be explained by seeing that larger sizes make the gates more resilient and prevent logic failures even in the presence of breakdown current. This causes the failure regions in Figure 6 to shrink, counteracting the tendency of larger areas to be susceptible to more failures.

7. Acknowledgments

This work was supported in part by the NSF under award CCR-0541367 and the SRC under contract 2007-TJ-1572. The second author thanks Prof. Montserrat Nafria for early discussions on this topic.

8. References

[1] J. H. Stathis. Physical and predictive models of ultrathin oxide reliability in CMOS devices and circuits. *IEEE Trans. Device and Mater. Rel.*, 1(1):43–59, March 2001.

[2] E. Y. Wu et al. CMOS scaling beyond the 100-nm node with silicon-dioxide-based gate dielectrics. *IBM J. Res. Dev.*, 46(2/3):287–298, March/May 2002.

[3] F. Crupi et al. 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.*, 3(1):8–13, March 2003.

[4] H. Wang et al. Impact of random soft oxide breakdown on SRAM energy/delay drift. *IEEE Trans. Device and Mater. Rel.*, 7(4):581–591, December 2007.

[5] R. Degraeve et al. Relation between breakdown mode and location in short-channel nMOSFETs and its impact on reliability specifications. *IEEE Trans. Device and Mater. Rel.*, 1(3):163–169, September 2001.

[6] Y. H. Lee et al. Prediction of logic product failure due to thin-gate oxide breakdown. In *Proc. IRPS*, pages 18–28, March 2006.

[7] K. Chopra et al. A statistical approach for full-chip gate-oxide reliability analysis. In *Proc. ICCAD*, pages 698–705, November 2008.

[8] R. Fernández et al. Gate oxide wear-out and breakdown effects on the performance of analog and digital circuits. *IEEE Trans. Electron Devices*, 55(4):997–1004, April 2008.

[9] B. Kaczer et al. Impact of MOSFET gate oxide breakdown on digital circuit operation and reliability. *IEEE Trans. Electron Devices*, 49(3):500–506, March 2002.

[10] B. Kaczer et al. Consistent model for short-channel nMOSFET after hard gate oxide breakdown. *IEEE Trans. Electron Devices*, 49(3):507–513, March 2002.

[11] Predictive Technology Model. `http://www.eas.asu.edu/~ptm/`.

[12] S. S. Sapatnekar et al. An exact solution to the transistor sizing problem for CMOS circuits using convex optimization. *IEEE Trans. Comput.-Aided Des.*, 12(11):1621–1634, November 1993.

[13] The MOSEK Optimization Software. `http://www.mosek.com/`.

Appendix

Proof of Theorem 1: Since failures of different logic cells are independent, the circuit-level FP at time t, $\Pr^{(\text{ckt})}_{\text{fail}}(t)$, is calculated as:

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

Here, $\Pr^{(i)}_{\text{fail}}(t)$ 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 $\Pr^{(i)}_{\text{BD}}(t)$, and second, the breakdown causes a logic failure, which is captured with the FP $\Pr^{(i)}_{(\text{fail}|\text{BD})}$ from Section 3.2. Substituting Equation (2) to above:

$$\Pr^{(\text{ckt})}_{\text{fail}}(t) = 1 - \prod_{i\in\text{NMOS}} \left(1 - \Pr^{(i)}_{(\text{fail}|\text{BD})}\left(1 - \exp\left(-\left(\frac{\gamma_i t}{\alpha}\right)^{\beta} a_i\right)\right)\right). \quad (17)$$

This equation gives the circuit FP, incorporating considerations related to the effective stressing 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 notation: denote $\Pr^{(\text{ckt})}_{\text{fail}}(t)$ by P_f, $\Pr^{(i)}_{(\text{fail}|\text{BD})}$ by p_i, and $(\frac{\gamma_i t}{\alpha})^{\beta} a_i$ by μ_i. Then, taking the logarithms of each side of (17):

$$\ln(1 - P_f) = \sum_{i\in\text{NMOS}} \ln\left(1 - p_i\left(1 - \exp\left(-\mu_i\right)\right)\right). \quad (18)$$

Using the first-order Taylor expansion, $\exp(-x) = 1 - x$ for $x = \mu_i$,

$$\ln(1 - P_f) \approx \sum_{i\in\text{NMOS}} \ln(1 - p_i\mu_i). \quad (19)$$

Using another first-order Taylor expansion, $\ln(1 - x) = -x$, $x = p_i\mu_i$, the approximation is further simplified as

$$\ln(1 - P_f) \approx -\sum_{i\in\text{NMOS}} p_i\mu_i. \quad (20)$$

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:

$$\Pr^{(\text{ckt})}_{\text{fail}}(t) = 1 - \exp\left(-\left(\frac{t}{\alpha}\right)^{\beta} \sum_{i\in\text{NMOS}} \Pr^{(i)}_{(\text{fail}|\text{BD})}\gamma_i^{\beta} a_i\right). \quad (21)$$

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 $O(x^2)$ are much smaller.

We can convert Equation (21) to the following form:

$$\begin{aligned} W &= \ln\left(-\ln\left(1 - \Pr^{(\text{ckt})}_{\text{fail}}(t)\right)\right) \\ &= \beta\ln\left(\frac{t}{\alpha}\right) + \ln\sum_{i\in\text{NMOS}} \Pr^{(i)}_{(\text{fail}|\text{BD})}\gamma_i^{\beta} a_i. \end{aligned} \quad (22)$$

□

[7] The concerned circuit failure is usually at the low end, e.g. $P_f < 0.1$. Due to the weakest-link property, the breakdown probability of each individual cell $\Pr^{(i)}_{\text{BD}}$ in a large circuit must be very small, which implies that μ_i is very small and must be far less than 1 (considering $\mu_i = 1$ means $\Pr^{(i)}_{\text{BD}} = 0.632$ for unit-size device). These approximations are validated by experimental results in Section 6.