# Analysis of System-Level Reliability Factors and Implications on Real-time Monitoring Methods for Oxide Breakdown Device Failures

Eric Karl, Dennis Sylvester, David Blaauw
*University of Michigan*
*{ekarl,dennis,blaauw}@eecs.umich.edu*

## Abstract

*Continued technology scaling exacerbates the incidence of degradation and failure in integrated circuits due to mechanisms such as oxide breakdown, negative bias temperature instability and electromigration. This work analyzes the impact of different factors on lifetime distributions for the oxide breakdown effect using a novel monte carlo approach based upon the percolation model and BSIM4. Results of the analysis of oxide failure distributions are used to explore real-time lifetime projection and the use of in-situ monitoring circuits. Under an ideal sensor assumption, the work shows that 500-1000 sensors would be needed to provide lifetime projections with error under 8-10%.*

## 1. Introduction

The continued, aggressive downscaling of dimensions in forthcoming CMOS technology generations [1] stands to increase the risk of significant reliability issues in integrated circuits. Conventional or constant electric field scaling law, in table 1, dictates that supply voltages must be reduced to maintain a constant electric field as the critical dimensions of CMOS transistors and wires shrink. However, in recent years, the voltage scaling trend of constant electric field scaling [2] has been violated to maintain the saturation current and other performance metrics.

Oxide breakdown (OBD) and the negative bias temperature instability effect (NBTI) are two reliability failure mechanisms for integrated circuits that are strongly impacted by electric fields in the gate oxide region. Both mechanisms degrade device structure through collisions between particles in the oxide lattice or the oxide-silicon interface region. Higher electric fields and temperatures lead to vastly different rates of degradation in MOSFET devices.

Before a reasonable approach to controlling reliability breakdowns and degradation can be proposed, there are important questions to answer about the nature of the problem. Is it typical to see outlying device failures or will failures occur steadily after the first failure? How will temperature, voltage, process variation and even circuit activity impact the probability of failure or degradation? Is there significant spatial correlation between failure events? This work attempts to address the aforementioned issues and determine the most effective methods for reducing the probability of failure or degradation through voltage, temperature or activity reduction. The conclusions from the analysis are used to explore the limitations of a real-time monitoring approach to understanding and controlling reliability issue. Assuming ideal sensors, confidence bounds for the error of real-time projection methods are presented as a function of sensor count.

Section 2-3 presents the modeling used to simulate the failure distributions for the oxide breakdown mechanism. Oxide breakdown is an ideal mechanism for this study since there are a variety of mature modeling approaches and it is a significant limiting factor on the scaling of device dimensions in future technologies. Section 4 will present results from simulation on the distribution of failure for the oxide breakdown mechanism. Section 5 details the implications on sensor-based monitoring techniques.

## 2. Oxide Breakdown Modeling

Oxide breakdown, or dielectric breakdown, is a degradation mechanism that results in a low-impedance path through an insulating or dielectric barrier. Failures related to this low-impedance path are typically manifest as abnormally high off-state leakage current, changes in circuit switching delay or even failure to switch in severe cases of degradation.

The percolation model, proposed by DeGraeve [3], treats oxide degradation as a series of traps or defects generated in the oxide layer. During operation, each electron passing a dielectric barrier has a small probability to enter a high-energy state to tunnel through the insulating layer and collide with particles in the lattice, possibly creating oxide traps or defects. Chaining paths of these oxide defects in the dielectric barrier reduce the energy level required for conduction through the layer, and therefore increase the probability that electrons will travel through the layer.

Defect generation of tunneling charges is the wear-out mechanism for thin dielectric films in the percolation model. When a critical defect density inside the oxide volume is reached, there is a high probability that a low-impedance defect path ultimately leads to uncontrolled current and oxide breakdown. The relationship between charge tunneling through the oxide and the defect density is expressed below in (1), where $N_{BD}$ is the defect density, $P_{DG}$ is the probability of defect generation, and $I_{tunnel}$ is the tunneling current, $V$ is the voltage across the oxide, and $T$ is the temperature [4].

$$N_{BD} \approx \int^{t} P_{DG}(V,T) I_{tunnel}(V,T) dt \qquad (1)$$

The percolation model places defects of a certain size into a 3-D oxide volume until a path of overlapping defects is created between the top and bottom planes. Repetitive simulation at a given dielectric thickness results in a probability density function for a chain of defects as a function of the defect density. Using the defect density, the probability of defect generation and the injected charge or tunneling current are used to calculate the time to formation of the defect chain, according to the relationship in Eq. (1).

The tunneling current through a gate oxide is calculated using BSIM4 model equations [5] in this work. In this work, published defect generation relationships from an IBM technology node are used in the simulations [4]. The empirically collected data is fit to an equation relating the voltage and temperature [6] of the oxide layer to the probability of defect generation. In (2), $V_{DD}$ is the stress voltage applied to the oxide and $T$ is the temperature of the oxide.

$$\begin{aligned} P_{DG} &= 4.9 \times 10^{-18} \cdot \exp\left(9.5946 \cdot V_{DD}\right) \cdot \\ &\quad \exp\left(3.85\left(\tfrac{1000}{T}\right)^2 - 29.679\left(\tfrac{1000}{T}\right) + 45.749\right) \end{aligned} \qquad (2)$$

## 3. Simulation Methodology

Using the oxide breakdown model detailed in the previous section, a monte carlo simulation methodology is utilized to generate a variety of oxide breakdown distributions with differing voltage, temperature, variation and stress time inputs. The diagram depicting the basic method used to generate the oxide breakdown distributions is outlined in Fig. 1.

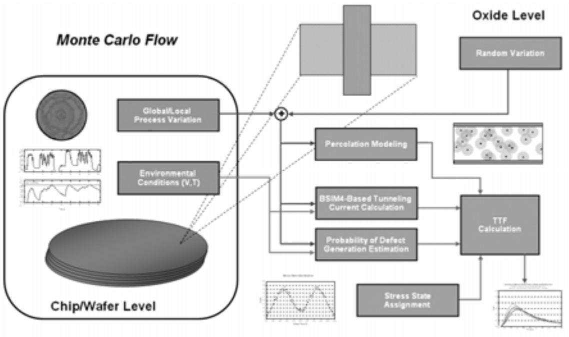
Monte Carlo Flow
Oxide Level
Random Variation
Global/Local Process Variation
Environmental Conditions (V,T)
Percolation Modeling
BSIM4-Based Tunneling Current Calculation
Probability of Defect Generation Estimation
TTF Calculation
Stress State Assignment
Chip/Wafer Level

Figure 1. Monte carlo simulation framework for generating oxide breakdown distribution.

The simulation method begins by generating a mapping of spatial variation in oxide length, width and thickness using a multi-level variation model. The variation model is a four-level tree model, with a global value at the top and random components on the lowest level. The values are normally distributed with parameter appropriate levels of variation listed in table 1. Voltage and temperature values are statically set for the simulations presented in this work.

Table 1. Parameter Values for Simulation Framework

| Symbol | Quantity | Value |
|---|---|---|
| $L_{drawn}$ | channel length | 130 nm |
| $V_{th0}$ | device threshold voltage | 250 mV |
| $VDD_{nom}$ | nominal supply voltage | 1.2 V |
| $T_{ox}$ | oxide thickness | 1.7-1.9 nm |
| $\sigma_{1\text{-}TOX}$ | global variation sigma for $T_{ox}$ | 0.036 nm (2.0%) |
| $\sigma_{2\text{-}TOX}$ | 2nd tier variation sigma for $T_{ox}$ | 0.009 nm (0.5%) |
| $\sigma_{3\text{-}TOX}$ | 3rd tier variation sigma for $T_{ox}$ | 0.009 nm (0.5%) |
| $\sigma_{R\text{-}TOX}$ | random variation sigma for $T_{ox}$ | 0.036 nm (2.0%) |
| $\sigma_{1\text{-}W/L}$ | global variation sigma for W/L | 4.0 nm / 2.6 nm |
| $\sigma_{2\text{-}W/L}$ | 2nd tier variation sigma for W/L | 1.0 nm / 0.65 nm |
| $\sigma_{3\text{-}W/L}$ | 3rd tier variation sigma for W/L | 1.0 nm / 0.65 nm |
| $\sigma_{R\text{-}W/L}$ | random variation sigma for W/L | 4.0 nm / 2.6 nm |

The variation of length, width and oxide thickness is used to initialize the 3-dimensional oxide container for the percolation placement simulation. The small variations influence the density required to create a continuous chain between the top and bottom of the oxide. The defect density needed to complete the chain is determined from the percolation and used to calculate the time-to-failure.

The variation information and the voltage/temperature data is used by the BSIM4 model

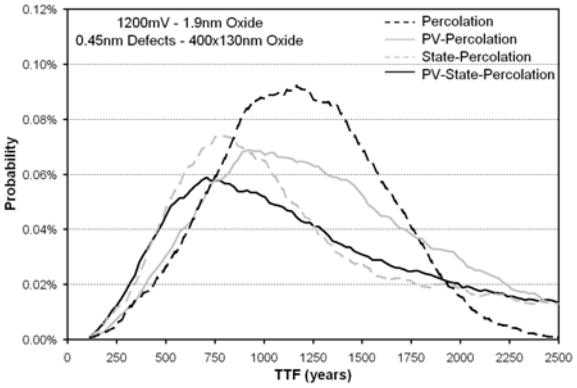

Figure 2. Oxide failure distributions with different simulation components included.

and the empirically fit probability of defect generation relationships to calculate the TTF of the oxide being simulated. The stress state, or the fraction of real time that the oxide is under max stress, is sampled from a simple bimodal distribution with peaks around 20% and 80%. This value directly modifies the "rate of injected charge" in a linear fashion, adding a realistic factor to spread the distribution of oxide breakdown that would be typical in a system not containing all oxides stressed continuously.

In the final step, the probability of defect generation, gate current and stress state are combined using the relationship in (1) to calculate the TTF for a single oxide in the simulated chip. The simulation can be used to characterize a sufficient number of oxides from one "chip-level" variation map and voltage/temperature profile. The entire process loops at the chip-level to develop distributions for different sets of process variation under the same parameters and voltage/temperature traces.

## 4. Failure Distribution Analysis

The simulation framework was used to generate a variety of oxide breakdown distributions at different oxide thicknesses, temperatures and voltages. An important goal for the analysis is to explore the impact of various inputs to the reliability model for oxide breakdown. To initially present the impact of process variation, state dependence and the inherent randomness of oxide breakdown, a simulation with fixed oxide dimensions, voltage and temperature is explored. The voltage is fixed to 1.2V, temperature at 350K and the oxide is 400nm x 130nm.

Graphically, in fig. 2, the lifetime distribution for a chip composed of 25,000 oxides is presented and the change in the distribution can be seen as the factors are added to the analysis one by one. Initially, the percolation model is used alone to observe the innate randomness of the process, and this results in a distribution that clearly matches a Weibull, which is used in most oxide breakdown projection methodologies. However, as the effect of process variation is added, the curve PV-Percolation in fig. 2 shows a distribution that has shifted to appear lognormal. The variation in the oxide thickness and the oxide size causes an exponential effect on the tunneling current and injected charge, which introduces this lognormal shape. The effect of state dependence, the fraction of lifetime that the oxide spends at high stress (when not at stress, degradation is assumed to be 0), spreads the distribution, pushing the peak of early failures to an earlier predicted year and smoothing the long tail on the right side. The trends in distribution shape shown in Fig. 2 determine the appropriate fitting functions for real-time lifetime projection in Section 5.

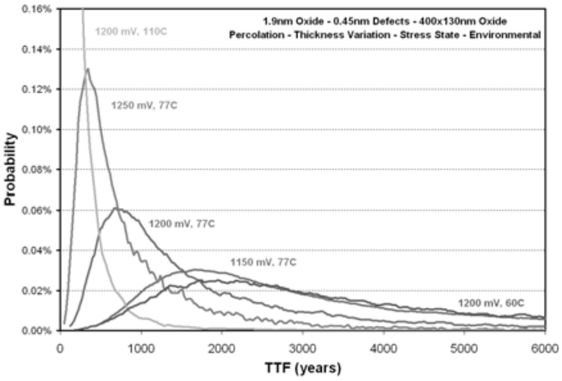

Figure 3. Failure distribution for 1.9nm oxide with conditions varying from 1.15-1.25V and 60-110C.

Figure 3 is a plot of the failure distribution of a 1.9nm oxide with 400x130nm dimensions considering percolation, process variation, stress state and varying environmental conditions. The effects of environmental conditions on the failure distribution of the oxides is enormous in this plot, particularly so for the wide range of temperatures (60-110C). The voltage range is selected to be on the order of a static offset in a regulator or a consistent small amount of noise in a power supply network. Even small voltage discretions lead to large changes in potential oxide lifetime due to exponential relationships in voltage and temperature. One of the greatest advantages of real-time monitoring of these effects is the ability to capture the impact of the environmental conditions on each die, which can be difficult to estimate for general-purpose components.

Figure 4 displays the 25th and 75th percentile values for the first failures of a simulation of 100 dies with 25,000 oxides on each die. A simulation size of 25,000 oxides provides reasonable simulation time and

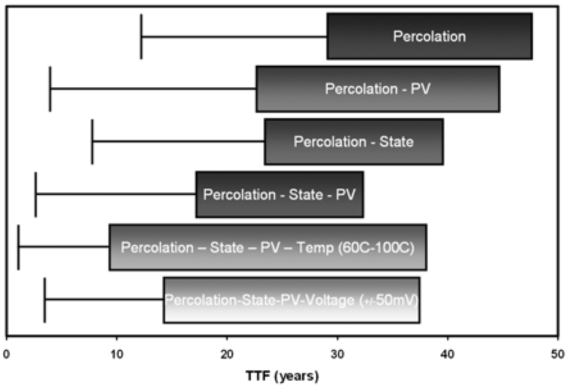

Figure 4. 25th to 75th percentile ranges for 100 die simulation of 1.9nm oxide first failures. The bar to the left of each range indicates the minimum TTF outlier. Each bar includes a different set of simulation components to demonstrate effects of process variation, state or temperature on failure.

exhibits the comparable results to a 250,000 oxide simulation. Starting with the basic percolation simulation at the top, the added effects typically reduce the TTF, for example the baseline percolation simulation minimum observed TTF of 12.17 years, becomes 1.01 years when considering process variation, state, and temperature effects.

Table 2. Spatial Correlation between Blocks

| Description | Expected | Simulated |
|---|---|---|
| 2nd failure in 3rd level block of first failure | 6.25% | 3.20% |
| 2nd failure in 2nd level block of first failure | 25.00% | 22.30% |
| 5 of first 5 failures share 2nd level block | 0.39% | 3.20% |
| 4 of first 5 failures share 2nd level block | 1.56% | 9.80% |
| 3 of first 5 failures share 2nd level block | 6.25% | 29.50% |

The 2nd level blocks are ¼ of the die area in the process variation model, and the 3rd level blocks are 1/16th of the die area in the model.

The effects of spatial correlation within the multi-level process variation model are analyzed in table 2. From the simulations of 100 dies consisting of 25,000 oxides, there are no signs of direct spatial correlation between the first observed failure and the second observed failure. However, when considering the first 5 or 10 failures, the data supports a moderate level of spatial correlation in large (25% die area) blocks. Monitoring circuits may be able to detect large areas that may be more susceptible to oxide failure, but to pinpoint a region for a second failure following an observed first failure is not likely.

The analysis of failure distributions using the simulation methodology answers many of the crucial questions posed in the introduction. There is an innate randomness to the oxide breakdown effect, and outlying failures are typical. Voltage and temperature have a dominant effect on predicted failure time and effects like state dependence and process variation alter the shape of the distribution from a pure weibull shape to nearly lognormal. These observations guide the exploration of the use of real-time monitoring in section 5.

## 5. Real-time Monitoring

Recent research into dynamic systems and reliability management has proposed the use of in-situ sensors to improve the inputs to model-based algorithms. The results of section 4 show that much can be gained from a reliability standpoint if you have some awareness of environmental conditions and the process variation of the die. Process variation and environmental conditions can be measure in real-time with known circuit techniques, yet even with knowledge of these values, a layer of modeling is needed to extrapolate the impact on reliability mechanisms. If a sensor could be designed to isolate and directly measure the degradation due to a particular mechanism, this layer of uncertain modeling could be eliminated. Based upon the analysis in Section 4, if an ideal sensor to detect the TTF for an oxide device under test could be designed, we aim to explore the bounds on accuracy and the requirements to gain some benefit from a real-time monitoring system.

Based upon the near-lognormal distribution shapes from section 4, least-squares method, a modified least-squares method and maximum likelihood estimators for lognormal distributions were used to fit samples of simulated distributions to obtain a prediction. The most accurate method for accurate prediction was least-squares fitting with the fit range censored to the earliest 30% of the sample set to ensure a good fit in the early failure range. Subsequent results on the quality of real-time monitoring approaches assumes the modified least-squares fitting method.

From the high degree of innate randomness in oxide breakdown failures, it is clear that multiple sensors will be needed to obtain any information when directly measuring the oxide degradation. To analyze the effects of the number of sensor samples needed to effectively predict the TTF for a die, a simulation of a single die was performed and the failure times for the die are sampled (assuming the samples are the available sensors) with different sensor counts, ranging from 35 sensors to 5000 sensors on the die. The error from the predicted value of TTF using the modified LS method is used as a figure of merit and the findings are listed in Table 3. For the die in question, the actual failure time is 4.841 years (1.8nm oxide). From this table, the sensor prediction approaches 10% average error around 1000 sensors. The clear point is that if direct measurement is intended, many sensors will be required and they will need to be very compact and accurate to provide a reasonable degree of accuracy.

Table 3. Sensor Count and Prediction Error
TTF of Die = 4.841 years

| Sensors | Abs. Mean Error | Min Error | Max Error |
|---|---|---|---|
| 35 | 3.7148 | -3.5437 | 18.0896 |
| 50 | 2.5499 | -3.3444 | 18.7818 |
| 100 | 1.4777 | -3.8320 | 6.8266 |
| 250 | 0.9150 | -2.3437 | 3.6080 |
| 500 | 0.7514 | -1.8663 | 3.1098 |
| 1000 | 0.4415 | -1.1737 | 2.3487 |
| 1500 | 0.3537 | -1.2094 | 1.0934 |
| 2000 | 0.3339 | -0.9953 | 1.3249 |
| 5000 | 0.1929 | -0.5549 | 0.6913 |

Figure 5 displays the 95% confidence bounds for 2 different dies using the modified LS method as sensor count increases. The worst case PVT line represents a typical corner estimate of the TTF for a chip from this process considering 85C, high voltage and pessimistic process variation. With low amounts of sensors, the 95% confidence bound can be worse than the corner estimate, yet with 500-1000 sensors, even a 95% confidence bound will result in a much better prediction than the corner model labeled worst-case PVT. A carefully designed real-time monitoring system can realize excellent improvements in TTF awareness (and the dynamic control schemes possible with this knowledge) over traditional corner-based reliability qualification.

Facing implementation of 500 or more sensors to directly monitor reliability mechanisms places strict constraints to realize a feasible system. The sensor design needs to be on the order of a standard cell macro block to ease placement and routing for such a large number of blocks. The accuracy requirement is high, since under the ideal assumption in this first look, large numbers of sensors are needed to reach an accurate prediction. In an era of billion transistor systems, 500-5000 sensors are a feasible budget to control the difficult problem of a priori reliability qualification.

## 5. Summary

This work presents a new method for oxide lifetime simulation that delivers lifetime distributions using a computationally reasonable approach. Analysis of the impact of process variation, state dependence and environmental conditions on the lifetime distribution of thin-film oxides is presented using the methodology. The conclusions from the analysis of oxide lifetime distributions are used to explore the feasibility of direct real-time measurement of a highly random breakdown mechanism in a system. Assuming an ideal sensor is available; 1000-2000 sensors throughout a chip can provide lifetime predictions with less than 10% average error. Current and future work on small, embedded process and reliability sensors is underway and future work on the application and accuracy of sensors implemented in silicon is planned.

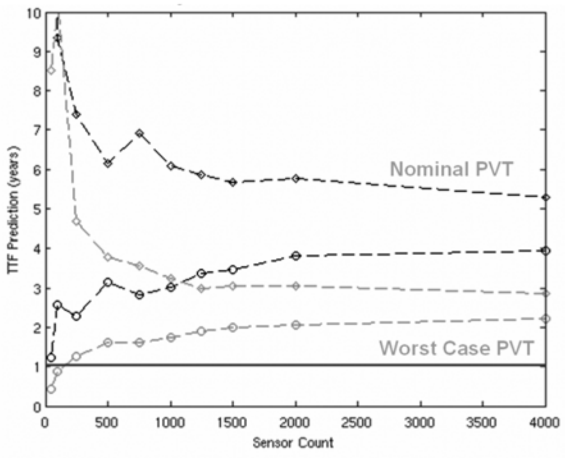

Figure 5. 95% confidence interval in TTF prediction as sensor count increases for 2 different dies.

## 6. References

[1] ITRS Reports, http://www.itrs.net/Links/2006Update/2006UpdateFinal.htm

[2] A. Chandrakasan, editor, *Design of High-Performance Microprocessor Circuits*. Piscataway, NJ: IEEE Press, 2001, pp. 27-28.

[3] R. Degraeve, et. al, “A consistent model for intrinsic breakdown in ultra-thin oxides,” *Intl. Electron Devices Meeting*, Dec. 1995, pp. 863-866.

[4] J. H. Stathis, “Physical and Predictive Models of Ultra Thin Oxide Reliability in CMOS Devices and Circuits,” *IEEE Trans. Device & Materials Reliability*, Vol. 1, Issue 1, March 2001, pp. 43-59.

[5] K.M. Cao, et. al., “BSIM4 Gate Leakage Model including Source-Drain Partition,” *IEDM Technical Digest*, 2000.

[6] E. Wu, D. L. Harmon, L-K Han, “Interrelationship of Voltage and Temperature Dependence of Oxide Breakdown for Ultrathin Oxides,” *IEEE Electron Device Letters*, Vol. 21, No. 7, July 2000.