A Statistical Approach for Full-Chip Gate-Oxide Reliability Analysis

Kaviraj Chopra, Cheng Zhuo, David Blaauw, Dennis Sylvester
Department of Electrical Engineering & Computer Science, University of Michigan, Ann Arbor, MI 48109
Email: {kaviraj, czhuo, blaauw, dennis}@eecs.umich.edu

***Abstract*— Gate oxide breakdown is a key factor limiting the useful lifetime of an integrated circuit. Unfortunately, the conventional approach for full chip oxide reliability analysis assumes a uniform oxide-thickness for all devices. In practice, however, gate-oxide thickness varies from die-to-die and within-die and as the precision of process control worsens an alternative reliability analysis approach is needed. In this work, we propose a statistical framework for chip level gate oxide reliability analysis while considering both die-to-die and within-die components of thickness variation. The thickness of each device is modeled as a distinct random variable and thus the full chip reliability estimation problem is defined on a huge sample space of several million devices. We observe that the full chip oxide reliability is independent of the relative location of the individual devices. This enables us to transform the problem such that the resulting representation can be expressed in terms of only two distinct random variables. Using this transformation we present a computationally efficient and accurate approach for estimating the full chip reliability while considering spatial correlations of gate-oxide thickness. We show that, compared to Monte Carlo simulation, the proposed method incurs an error of only 1∼6% while improving the runtime by around three orders.**

I. Introduction

Semiconductor reliability and manufacturing variability have become key issues as device critical dimensions shrink and integration complexity continues to grow at a rapid pace. For assessing product reliability, it is important to quantify the reliability of gate-oxide which is its ability to retain its dielectric properties while being subjected to high electric fields. Aggressive oxide-thickness scaling has led to huge vertical electric fields in metal oxide semiconductor devices that result in high direct tunneling gate oxide leakage current. The gate oxide leakage current creates defects such as electron traps, interface states, holes traps, etc., in the gate dielectric. These defects gradually build up in the oxide until a critical defect density is reached when the oxide destructively breaks down leading to a large increase in gate oxide conductance and functional failure of the product.

Over the last 30 years, numerous publications have focused on understanding and modeling the mechanisms leading to defect generation and breakdown in individual devices [1]. Some researchers have initiated an effort to understand the oxide breakdown mechanisms of simple circuits [2]. Recently, a product level approach performing oxide breakdown analysis on full-chip was proposed in [3]. In most of the existing approaches, simple test structures such as discrete devices or capacitors are used to characterize the oxide breakdown mechanism for a specific manufacturing process. These discrete device characterization results are then extrapolated to deduce a model for the full-chip oxide reliability which is later calibrated using lifetime tests of sample product.

However, a major concern with prior approaches is that they assume a uniform oxide-thickness for all devices on every chip. In practice, the non-uniformity in temperature and pressure during the gate-oxidation process leads to within-die and die-to-die variations in gate-oxide-thickness. For a given supply voltage and operating temperature, the reliability of oxide is an exponential function of its thickness and its sensitivity to thickness variations increases for thinner oxides [4]. Therefore, in previous approaches, it was imperative to consider a uniform minimum oxide-thickness across all devices on a chip and across all chipss for a conservative worst-case analysis. This may lead to significantly pessimistic estimates of the overall oxide breakdown reliability of the product. Furthermore, oxide reliability is one of the key factors that sets constraints on the operating supply voltage and temperature of the chip. Any pessimism in oxide reliability analysis limits the maximum operating voltage and thus the maximum achievable chip-performance. In order to find consistent supply voltage and operating temperature limits, it is therefore critical to quantify the product oxide breakdown strength more effectively.

The goal of this work is to develop a new chip level gate oxide breakdown analysis while considering not only the chip-to-chip and within-chip variations but also spatial correlations of gate-oxide thickness. If the thickness of each device is modeled as a distinct random variable, then the full chip reliability estimation problem is defined on a huge sample space of several million devices. By noting that the reliability for a sample device with a given oxide-thickness itself is a random function, the design time full-chip reliability estimation problem turns out to be a multi-dimensional nested stochastic process. Apparently a straightforward Monte Carlo approach is extremely expensive in both execution time and memory, as we need to perform nested Monte Carlo analysis on the sample spaces for different chips and different devices across each chip as well as the sample space of oxide breakdown of each device. The challenge here is how to reduce the tremendous number of random variables and then achieve a low space/time complexity. To address this issue, we propose a statistical approach that can project millions of random variables to only two distinct random variables of mean and variance of chip oxide-thickness distribution. Then, with some judicious approximations, we can use this projection to compute the closed-form reliability function for each chip. Finally, a computationally efficient and accurate statistical framework is developed to estimate the full chip reliability across the ensemble of chips. To our knowledge, this is the first attempt to perform the chip-level oxide reliability analysis while modeling the components of oxide variation and correlations.

Apart from manufacturing variations in oxide-thickness, the operational life of a part depends on variations in temperature, modes of operation, executed instructions, supply voltage, lifetime wear out, etc. However, the uncertainty of the environmental conditions of a part can not be modeled statistically as the chip is expected to function reliably throughout its operational lifetime even in a worst-case operating environments for all specified applications. Therefore, unlike manufacturing variability, our approach considers the worst-case operating temperature and supply voltage to ensure a correct operation through out the entire life time of the part for any application profile. Experimental results show that, compared to

978-1-4244-2820-5/08/$25.00 ©2008 IEEE

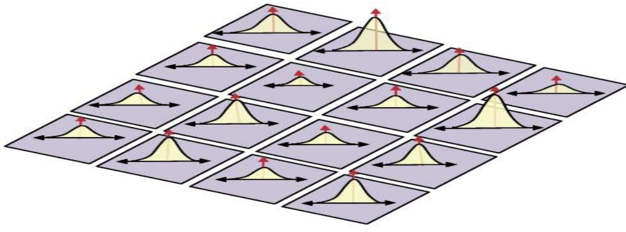

Fig. 1. A grid based spatial correlation model

Monte Carlo simulation, the proposed method incurs an error of only 1∼6%, while improving the runtime by around three orders.

The rest of the paper is organized as follows: in Section II we describe the modeling of thickness variation. In Section III, we discuss the oxide breakdown model and formulate the oxide reliability analysis problem. Section IV explains the proposed methodology for estimating the full-chip oxide breakdown reliability. Simulation results illustrating the efficacy of the proposed approach are given in Section V. We conclude the paper in Section VI.

II. Modeling Thickness Variation

The oxide-thickness variation can be classified based on the spatial scale over which it manifests. Due to long range shifts in oxidation temperature and pressure that occur from lot-to-lot, wafer-to-wafer, reticle-to-reticle, and across a reticle, all the devices on the same chip observe some common amount of fluctuation in oxide-thickness. This chip-to-chip component of variations is referred to as *global* or *inter-chip variation*. Several factors gradually affect temperature from one location to another within a chip ($e.g.$, the emissivity variations resulting from location of chip on the wafer). Such variations tend to affect all devices that are placed close to each other in a similar manner. Therefore, closely spaced devices are more likely to have similar oxide-thickness than those placed far apart. The component of variation that exhibits such spatial dependence is known as *spatially correlated variation*. For accurate statistical analysis, it is necessary to capture the dependence between the global and spatial component of thickness variations. The residual variation resulting from certain local device scaling effects such as different surface orientations, stress conditions as well as poly-Si intrusion from the gate electrodes is referred to as *independent variation*.

To exactly model spatial correlation between the oxide-thickness of two devices, a separate random variable is required for each device. However, the correlation between two devices is generally a slow monotonically decreasing function of their separation. Therefore, simplified correlation structures using a grid model [5] or quad-tree model [6] have been proposed in the literature. In this work, we discuss the proposed approach using the grid based model. In this model, the spatial component of oxide-thickness variation is modeled using n random variables, each representing the spatial component of variation in one of the n grids (see Figure 1), and a covariance matrix of size $n \times n$ representing the spatial correlations among the grids. The covariance matrix could be determined from measurement data extracted from manufactured wafers using the method given in [7]. To simplify the correlation structure, this set of correlated random variables is mapped to another set of mutually independent random variables with zero mean and unit variance using the principal components of the original set. The original random variables are then expressed as a linear combination of the principal components. These principal components can be obtained by performing an eigenvalue decomposition of the correlation matrix. This representation of the correlation is expressed in a so-called canonical form [5], [8], [9], where oxide-thickness x_i of any device in i^{th} grid is given by

$$x_i = \lambda_{i,0} + \sum_{j=1}^{n} \lambda_{i,j} z_j + \lambda_r \epsilon, \tag{1}$$

where $\lambda_{i,0}$ is the mean or nominal value of oxide-thickness in i^{th} grid , z_j represents the n independent random variables used to express the spatially correlated device parameter variations, ϵ is a distinct random variable for each device that represents the residual independent variation, and the coefficients $\lambda_{i,j}$'s represent the sensitivity of thickness variation in i^{th} principal component for every j^{th} the random variable.

III. Reliability Model and Problem Formulation

The gate oxide degradation depends on the oxide-thickness, voltage, and temperature. There are many oxide breakdown models in the literature that attempt to explain the dependence on these factors. A widely accepted model is the anode hole injection model [10]. According to this model, injected electrons generate holes at the anode that can tunnel back into the oxide. Intrinsic breakdown occurs when a critical hole fluence is reached, creating a continuous conducting path across the oxide. A second model, known as electron trap density model, has been suggested, which claims that a critical density of electron traps generated during stress is required to trigger oxide breakdown [11]. Both models of oxide breakdown mechanisms note that the defect generation is a non-deterministic process. As a result the oxide breakdown time is inherently a statistically distributed quantity. Thus the oxide breakdown time is modeled as a random variable typically characterized by a Weibull probability distribution function, given by [4], [12]

$$F(t) = 1 - e^{-a(\frac{t}{\alpha})^{\beta}}, \tag{2}$$

where F is the cumulative distribution function (CDF) of time-to-breakdown t, a is the device area normalized with respect to the minimum device area, α and β are the scale and shape parameters of the Weibull distribution. The scale parameter α represents the characteristic life which is the time where 63.2% of samples fail, whereas the shape parameter β is a function of critical defect density. The critical defect density depends on device oxide-thickness, the oxide field and temperature. For a given temperature and stress voltage, it has been shown that the slope parameter of the Weibull distribution varies linearly with oxide-thickness [13]. Thus if x denotes the oxide-thickness, we have

$$F(t) = 1 - e^{-a(\frac{t}{\alpha})^{bx}}, \tag{3}$$

where b is a constant for given worst-case temperature and supply voltage. Another major factor that affects the oxide lifetime is the oxide breakdown failure criterion. A commonly used failure criterion is soft breakdown (SBD) which is characterized by a small increase in gate leakage. In practice, however, after SBD the gate leakage current monotonically increases with time eventually leading to a hard breakdown [14]. The time between oxide SBD and hard breakdown is a function of the gate area, oxide quality, and the bias conditions. For the purpose of this work, we limit our analysis to determining the initiation of soft breakdown and use this as our failure criteria since SBD is typically followed rapidly by hard breakdown.

A chip is considered to have failed as soon as breakdown occurs for any device on the chip. We are interested in finding the reliable lifetime of the chip for which none of the devices fail. For such weakest link problems, it is more convenient to use an alternate

representation known as reliability function $R(t)$ or survivor function, given by

$$R(t) = 1 - F(t) = P(T > t) = \int_t^\infty f(s)ds, \quad (4)$$

where $f(s)$ is the probability distribution function (PDF) of oxide breakdown of an individual device. The reliability function is complimentary to the cumulative distribution function (CDF) $F(t)$, taking the value 1 at $t = 0$ and tending to 0 as t tends to infinity. Simply stated, a reliability function is the probability that a device (chip) does not fail by time t. Due to manufacturing variations the thickness of gate oxide is also a non-deterministic parameter at design time. Thus the reliability function of a device can be interpreted as its conditional reliability function for a given oxide-thickness. For an ith device having x_i oxide-thickness the conditional reliability function can be given as

$$R_i(t|x_i) = P(t > \mathbf{t}|x_i) = \int_t^\infty f(s|x_i)ds. \quad (5)$$

Due to the spatial component of oxide-thickness variation, the oxide-thicknesses of any two devices on a chip are correlated with each other. Therefore, in general, their respective reliability functions being functions of oxide-thickness are also correlated with each other. However, if the oxide-thicknesses are known apriori then the defect generation mechanism in one device is independent of any other device on the chip for constant worst-case voltage and temperature. Thus for a particular chip, if the thicknesses of all devices are known then any device fails independently of all other devices. Furthermore the reliability function of the chip $R_c(t)$ requires that all devices on the chip are functioning reliably, therefore, $R_c(t)$ is given by the product of reliability functions of all individual devices:

$$R_c(t|\mathbf{x}) = \prod_{i=1}^{m} R_i(t|x_i), \quad (6)$$

where $\mathbf{x}$ represent the vector of oxide-thickness $(x_1, \ldots, x_m)$ and m is the total number of devices on the chip.

In the traditional analysis, where all oxide-thicknesses are supposed to have a single, worst-case value, the product in Equation (6) is taken across a large set of identical reliability functions and can be analytically solved with a low complexity. However, the key point in our analysis is that, at design time, each oxide-thickness x_i is itself a random variable. Furthermore, these random variables are correlated across the chip. If the oxide-thickness of all devices is characterized by their joint PDF $f(x_1, \ldots, x_m)$ then the overall reliability function of the entire ensemble of all manufactured chips can be given by

$$R_c(t) = \int_0^\infty \ldots \int_0^\infty \prod_{i=1}^{m} R_i(t|x_i) f(x_1, \ldots, x_m) dx_1 \ldots dx_m. \quad (7)$$

Due to the huge dimensionality of the above integral, a straight forward numerical evaluation of the above integral is computationally impractical for full chip analysis. Using judicious approximations we develop a computationally efficient approach to address this problem in the next Section. In Section 5 we then present results that demonstrate the accuracy and efficiency of the proposed solution.

IV. Full-Chip Gate-Oxide Reliability Analysis

The proposed approach for efficiently estimating the overall reliability function $R_c(t)$ is discussed in a bottom up manner. We first present expressions for finding the conditional reliability function of a device. Using this expression, the conditional reliability function of a particular chip can be found given the oxide-thickness of all devices on it. We observe that although the overall reliability functions depends on the spatial and global correlation in oxide-thickness variation, however, it is independent of the relative location of two devices on the chip. Hence, for a given chip, we can sum together all oxides of equal thickness and generate a frequency distribution histogram of the oxide-thickness. As the oxide-thickness variation of all the individual devices is modeled as a normal random variable and there are a large number of devices on a chip, we approximate this frequency distribution across a given chip with a normal distribution function. Henceforth, we will refer to this distribution function as the chip oxide-thickness distribution (COD). The COD allows us to compactly represent the oxide-thickness of all device on a given chip with just two parameters - the mean and the variance of the underlying normal distribution function. In sub-section IV-A, we will present how a closed form expression for the chip reliability function can be found for a given COD. Finally, we discuss how to compute the overall reliability function across the entire ensemble of all manufactured chips. As COD varies from chip-to-chip, its mean and variance are in fact random variables over the sample space of all manufactured chips. Hence, several million multi-variate oxide-thickness distribution function for each device on the chip can be compactly modeled with just two random variables. In sub-section IV-B, we will discuss how these two random variables can be derived from the oxide-thickness process variation model given in Equation (1) and thus the overall reliability function can be computed from it.

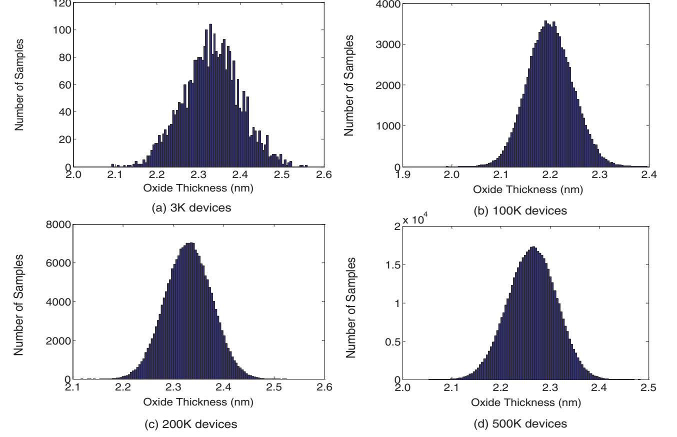

Fig. 2. Histogram of oxide-thickness for a sample chip with (a) 3K devices (b) 100K devices (c) 200K devices (d) 500K devices

A. Reliability Function of One Chip

Using the definition of the reliability function and the oxide breakdown time model of an individual device (Equation (3)), the conditional reliability function of an ith device on a chip having oxide-thickness x_i is given by

$$R_i(t|x_i) = e^{-a_i (\frac{t}{\alpha})^{b x_i}}. \quad (8)$$

As explained in Section III, if the oxide thickens of all devices on a chip is known then the reliability function of every device is independent of each other. Thus the reliability function of a chip is the product of the individual reliability function of all devices. Considering each device on the chip $x = (x_1, x_2, \ldots, x_m)$ and their respective area a_i, the conditional reliability of the chip is given by:

$$R_c(t|\mathbf{x}) = \prod_{i=1}^{m} R_i(t|x_i) = e^{-\sum_{i=1}^{m} a_i (\frac{t}{\alpha})^{b x_i}}. \quad (9)$$

There can be several million devices on a chip, therefore, it is impractical to evaluate the above exponent. In order to efficiently evaluate the overall reliability across all chips, we need to reduce the dimensionality of the above exponent. To achieve this, we represent the set of devices and their individual oxide-thicknesses by a COD for a particular chip. This chip-specific COD shows how many devices correspond to a particular oxide-thickness. For the sake of understanding, we discretize this oxide-thickness distribution into k discrete intervals assuming a truncated distribution. It can be seen that when we make this transformation the area of the devices with identical thickness can be summed together. Let $\overline{x}_i$ denote the midpoint of the ith discrete interval and $\overline{a}_i$ be the total area of all devices having thickness $\overline{x}_i$. By applying this transformation, the above expression for $R_c(t|\mathbf{x})$ can be rewritten as

$$R_c(t|\mathbf{x}) = e^{-\sum_{i=1}^{k} \overline{a}_i (\frac{t}{\alpha})^{b\overline{x}_i}}. \quad (10)$$

By making such a transformation the dimensionality of $R_c(t|\mathbf{x})$ can be reduced from number of devices m to the number of discrete intervals k. If we normalize the exponent with total area the above expression gives

$$R_c(t|\mathbf{x}) = e^{-A\sum_{i=1}^{k} p_i (\frac{t}{\alpha})^{b\overline{x}_i}}, \quad (11)$$

where $p_i = \overline{a}_i/A$ represents the probability of observing an oxide-thickness x_i on a particular sample chip. Thus the thickness of all devices on a particular sample chip can be compactly characterized by a COD.

As discussed in Section II, the thickness variation of an individual device across the ensemble of chips is modeled as a gaussian random variable. Although it is not guaranteed that the COD across a particular sample chip will also be a gaussian distribution, we find that assuming gaussian distribution for the oxide-thickness is a reasonable approximation to its exact distribution, due to the large number of devices on the chip. Numerical experiments further confirmed our assumption. Figure 2 illustrates the histogram of oxide-thickness for sample chips with different number of devices, ranging from 3K to 500K. It is clear that we get non-gaussian distribution for the chip with a small number of devices, $e.g.$ Figure 2(a), whereas the samples with a large number of devices show distinctly gaussian-like curves, $e.g.$ Figure 2(b)-(d). Thus we can formulate Equation (11) as:

$$R_c(t|u,v) = e^{-A\int_{-\infty}^{\infty} \phi(\frac{x-u}{\sqrt{v}})(\frac{t}{\alpha})^{bx}dx}, \quad (12)$$

where $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2}$.

In the above Equation (12), the integral in the exponent can be evaluated by making the substitution $\frac{t}{\alpha} = e^{\gamma}$,

$$\begin{aligned} &\int_{-\infty}^{\infty} \phi(\frac{x-u}{\sqrt{v}})(\frac{t}{\alpha})^{bx}dx = \int_{-\infty}^{\infty} \phi(\frac{x-u}{\sqrt{v}})e^{\gamma bx}dx \\ &= -\frac{1}{2}e^{\gamma bu+\gamma^2 b^2 v/2} erf(\frac{-x+u+\gamma bv}{\sqrt{2v}})|_{-\infty}^{\infty} \\ &= e^{\gamma bu+\gamma^2 b^2 v/2}. \end{aligned} \quad (13)$$

Thus for a given COD $\phi(\frac{x-u}{\sqrt{v}})$, the conditional reliability function of a chip can be computed by the following closed form expression:

$$R_c(t|u,v) = e^{-Ae^{\ln(\frac{t}{\alpha})bu+(\ln(\frac{t}{\alpha}))^2 b^2 v/2}}. \quad (14)$$

Hence, the multidimensional exponent in Equation (9) can now be compactly represented using a closed form analytical function of COD parameters u and v.

B. Overall Reliability Function

In this section, we will discuss how the overall reliability function can be found by enumerating the conditional distribution function derived in the previous section across the ensemble of all chips. As shown in Figure 3, each sample chip results in a different COD $\phi(\frac{x-u}{\sqrt{v}})$, therefore, the entire ensemble of all chips can be represented with set of CODs for all manufactured chips. Now each such COD is characterized by their respective mean u and variance v. Therefore, the oxide-thickness distribution of all devices across all manufactured chips can be represented by two random variables $\mathbf{u}$ and $\mathbf{v}$. Let $f_{\mathbf{uv}}(u,v)$ denote the joint probability distribution function of $\mathbf{u}$ and $\mathbf{v}$. For computing the overall reliability function, we need to integrate the above expression of reliability function of each chip over the joint probability distribution function $f_{\mathbf{uv}}(u,v)$ of random variables $\mathbf{u}$ and $\mathbf{v}$:

$$R_c(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} R_c(t|u,v) f_{\mathbf{uv}}(u,v)du\ dv. \quad (15)$$

For a particular chip, the mean u and and variance v of the oxide distribution can be estimated by calculating the unbiased statistical COD mean and variance of the oxide-thickness values observed across the chip. Likewise, the random variables $\mathbf{u}$ and $\mathbf{v}$ can be found in terms of the thickness variation model discussed in Equation (1). Using the oxide-thickness variation model given in Equation (1), $\mathbf{u}$ can be expressed as

$$\mathbf{u} = \frac{1}{m}\sum_{i=1}^{m} x_i = u_0 + \sum_{i=1}^{n} u_i z_i + u_{p+1}\epsilon, \quad (16)$$

where

$$\begin{aligned} u_j &= \frac{1}{m}\sum_{i=1}^{m} \lambda_{i,j} \quad \forall j = 0 \ldots n \\ u_{p+1} &= \frac{1}{m}\sqrt{\sum_{i=1}^{m} \lambda_r^2} = \frac{\lambda_r}{\sqrt{m}} \end{aligned}$$

The coefficient u_0 is the nominal value of $\mathbf{u}$, whereas coefficient u_i is the sensitivity to the ith principal component. It is evident that the sensitivity of the independent random component u_{p+1} tends to zero for a large number of devices and thus can be safely neglected for a typical industrial chip.

Similarly the expression for $\mathbf{v}$, the variance of the oxide distribution across the ensemble of all chips, in terms of oxide variation model (Equation (1)) can be given as follows:

$$\mathbf{v} = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \mathbf{u})^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i^2 - \mathbf{u}^2). \quad (17)$$

Again the above expression can be expressed in terms of principal components as follows:

$$\mathbf{v} = v_0 + \sum_{i=1}^{n}\sum_{j=1}^{n} v_{i,j} z_i z_j, \quad (18)$$

where

$$v_o = \lambda_r^2 \quad \text{and} \quad v_{i,j} = \frac{1}{n-1}\sum_{k=1}^{n}(\lambda_{k,i} - u_i)(\lambda_{k,j} + u_j)$$

In this manner, we can express the distributions of $\mathbf{u}$ and $\mathbf{v}$ in terms of a given process variation model. Note that random variable $\mathbf{u}$ is the sum of normal random variables so it is also a normal random variable, however, the COD variance v is not a normal random

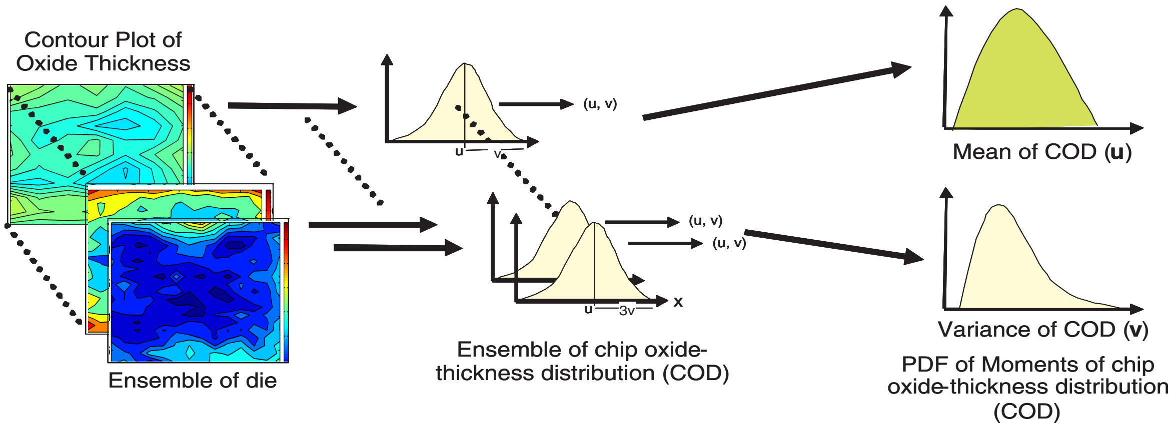

Fig. 3. Compact representation of oxide-thickness variation of ensemble of all chips

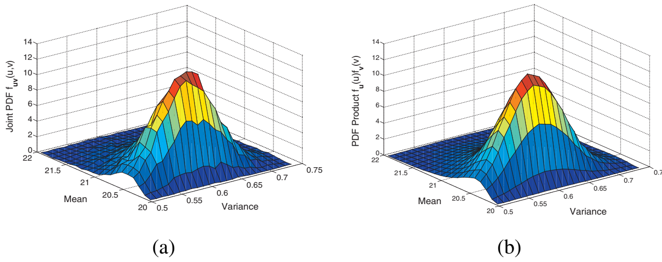

Fig. 4. (a) Joint PDF $f_{\mathbf{uv}}(u, v)$ (b) PDF product $f_{\mathbf{u}}(u)f_{\mathbf{v}}(v)$

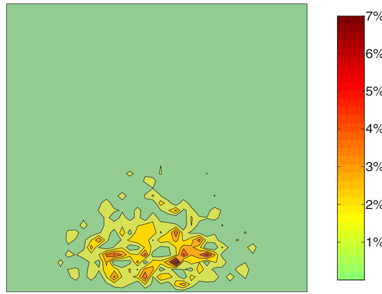

Fig. 5. Contour of the error between joint-PDF $f_{\mathbf{uv}}(u, v)$ and PDF product $f_{\mathbf{u}}(u)f_{\mathbf{v}}(v)$

variable. It can be shown as follows that $\mathbf{u}$ and $\mathbf{v}$ are uncorrelated (*i.e.*, $E[\mathbf{uv}] = E[\mathbf{u}]E[\mathbf{v}]$, where $E[\cdot]$ denotes the expectation):

$$E[\mathbf{uv}] = E[(u_0 + \sum_{i=1}^{n} u_i z_i + u_{p+1}\epsilon)(v_0 + \sum_{i=1}^{n}\sum_{j=1}^{n} v_{i,j} z_i z_j)]$$

By constructing each principal component z_i as an independent standard normal random variable, we have $E[z_i] = E[z_i^2 z_j] = E[z_i z_j^2] = E[z_i^3] = 0$ and $E[z_i^2] = 1$ for diffent i and j. Likewise $E[\epsilon] = E[\epsilon^2 z_j] = E[z_i \epsilon^2] = E[\epsilon^3] = 0$ and $E[\epsilon^2] = 1$. Thus the above expression can be simplified and given by

$$E[\mathbf{uv}] = u_0 v_0 + \sum_{i=1}^{n} u_0 v_{i,i} = E[\mathbf{u}]E[\mathbf{v}]. \tag{19}$$

For two normal random variables to be independent, it is sufficient to show that they are uncorrelated, but in general this is not the case for non-gaussian random variables. The sample variance $\mathbf{v}$ is not a normal random variable and has the distribution of quadratic forms in normal variables [15], [16]. However, with numerical experiments we find that the dependence between $\mathbf{u}$ and $\mathbf{v}$ is weak. As a result, it is reasonable to assume $\mathbf{u}$ and $\mathbf{v}$ as independent variables, which allows us to express the joint-PDF (JPDF) in terms of their marginal distributions $f_{\mathbf{u}}(u)$ and $f_{\mathbf{v}}(v)$. Figure 4 illustrates the JPDF of $f_{\mathbf{uv}}(u, v)$ and the product of the marginal distributions, $f_{\mathbf{u}}(u)f_{\mathbf{v}}(v)$, generated by Monte Carlo simulations. It is qualitatively evident from the figure that there does not exist significant dependence between $\mathbf{u}$ and $\mathbf{v}$. Furthermore, figure 5 depicts the contour of the error between JPDF $f_{\mathbf{uv}}(u, v)$ and PDF product $f_{\mathbf{u}}(u)f_{\mathbf{v}}(v)$ normalized with respect to the peak probability on the JPDF. It is noted that the maximal error is around 7% in a very small region whereas most errors are almost negligible. Thus, this independence approximation between $\mathbf{u}$ and $\mathbf{v}$ can give us a reasonably accurate estimate of oxide variation with a significantly simpler approach. In other words, the approximation enables us to enumerate the individual reliability distribution functions of each chip by simply integrating the marginal distributions $f_{\mathbf{u}}(u)$ and $f_{\mathbf{v}}(v)$:

$$R_c(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} R_c(t|u, v) f_{\mathbf{u}}(u) f_{\mathbf{v}}(v) du\, dv. \tag{20}$$

Now the COD mean $\mathbf{u}$ is a sum of normal random variables, therefore $f_{\mathbf{u}}(u)$ can be characterized by distribution of a normal random variable and can be analytically computed. However, COD variance v is a quadratic expression of normal random variables. Such an expression is commonly found in several multi-variate statistics application and is referred to as quadratic normal form. In statistics literature [15], several techniques have been proposed to accurately estimate the distribution function of quadratic normal form. In this work, we implemented a computationally efficient method given in [16] to estimate the distribution of $f_{\mathbf{v}}(v)$ using a chi-square approximation.

$$\mathbf{v} \sim v_0 + \hat{a}\chi^2_{\hat{b}}, \tag{21}$$

where $\hat{a} = \sum_{i=1}^{n}\sum_{j=1}^{n} v_{i,j}^2 / \sum_i v_{i,i}$ and $\hat{b} = (\sum_i v_{i,i})^2 / \sum_{i=1}^{n}\sum_{j=1}^{n} v_{i,j}^2$.

In figure 6, we compare the CDF of the distribution of quadratic normal form of $\mathbf{v}$ by Monte Carlo simulation and its chi-square approximation for the a chip with 100K devices. It is apparent that the computationally efficient chi-square representation is in good agreement with the Monte Carlo simulation result.

In this manner, the marginal distributions $f_{\mathbf{u}}(u)$ and $f_{\mathbf{v}}(v)$ of $\mathbf{u}$ and $\mathbf{v}$ can be analytically found for the given process variation model of oxide variation. Using $f_{\mathbf{u}}(u)$ and $f_{\mathbf{v}}(v)$, the overall reliability distribution function can be computed by evaluating a single two-dimensional numerical integration given in Equation (20).

TABLE I
ACCURACY COMPARISON BETWEEN PROPOSED APPROACH AND MC SIMULATIONS FOR DIFFERENT CORRELATION DISTANCE

circuit		Lifetime Estimation Error w.r.t. MC simulations					
		Rho-distance = 0.25		Rho-distance = 0.5		Rho-distance = 0.75	
name	No. devices	1/million	10/million	1/million	10/million	1/million	10/million
A	50K	4.30%	2.51%	4.23%	2.24%	4.59%	1.52%
B	80K	5.26%	3.58%	5.07%	3.25%	4.93%	2.28%
C	100K	6.35%	2.32%	5.62%	2.04%	4.57%	3.00%
D	200K	5.61%	3.30%	5.40%	2.57%	4.66%	1.89%
E	500K	5.72%	3.53%	5.01%	2.83%	4.86%	2.72%

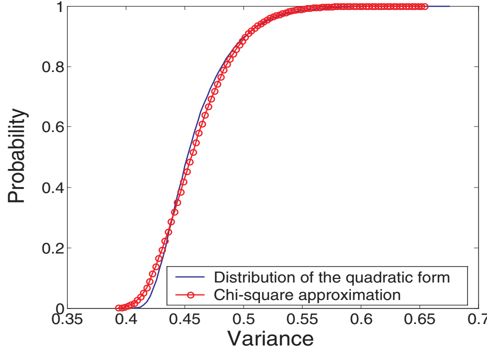

Fig. 6. Curves for the distribution of the quadratic form and its χ^2 approximation

C. Overall Algorithm and Complexity

The overall algorithm of the proposed approach is summarized as shown in figure 7. Given the principal components as well as the oxide thickness variations profile, we can compute the coefficients u_i and $v_{i,j}$ for COD mean and variance using Equations (16) and (18). Then we divide the integration domain for Equation (20) to $l \times l$ sub-domains. Since the JPDF rapidly decreases to 0 beyond a narrow domain, as illustrated in figure 4, l=10 is already a reasonable number for accurate integral sum evaluation, which is further confirmed by the experimental results in Section 5. Once sample point pair in each sub-domain is obtained, we can compute analytically the reliability for one chip. Finally the overall reliability is evaluated by using the integral sum.

Since PCA is a pre-processing step, we do not include it in the complexity analysis as it is performed only once and can be shared with other statistical analysis tools. As discussed in figure 7, steps 5-6 can be analytically computed using Equations (14), (16) and (21) and therefore will not limit the performance. Thus, the overall complexity is $O(n^2 + l^2)$, where n is the number of principal components and l^2 is the number of sub-domains for integration. By noting that unlike the straightforward approach, the computation complexity is independent of the total number of devices on the chip, it is therefore extremely computationally efficient compared with Monte Carlo analysis, whose complexity heavily depends on the number of devices.

V. SIMULATION RESULTS

A simple simulation methodology for estimating the critical defect density required for triggering a dielectric breakdown in an ultra thin oxide was originally developed by Degraeve in [13]. Using this methodology and the published defect generation relationships from a 130nm IBM technology node given in [12], the oxide breakdown reliability function was found for a set of characteristic devices differing in area and thickness. The technology dependent parameter of the oxide reliability function model given in Equation (3) was estimated by fitting it to the simulation results. In practice, such a model can also be characterized from real oxide breakdown distributions measured from test capacitors or discrete devices for the required process and technology. The proposed statistical approach was implemented with MATLAB.

Procedure: *Full-Chip Gate-Oxide Reliability Analysis*

Input: Number of devices, principal components using the variation model in Equation (1), chip-to-chip and within-chip variations profile, spatial correlation profile.

Output: Gate-oxide reliability.

1: Compute the coefficients u_i and $v_{i,j}$ for COD mean and variance with Equations (16) and (18);
2: Divide the integration domain for COD mean and variance to $l \times l$ sub-domains;
3: Compute the sample point pairs (u_i, v_j) for each sub-domains;
4: **For** each sample point pair **do**
5: Analytically compute the reliability for one chip using Equation (14);
6: Evaluate the PDF product $f_{\mathbf{u}}(u)f_{\mathbf{v}}(v)$ for the sample point;
7: **end for**
8: Compute the overall reliability with the integral sum;

Fig. 7. Algorithm for full-chip gate-oxide reliability analysis

In the first set of experiments, a set of 5 designs was considered to determine the accuracy of the proposed methodology. The overall 3σ of oxide-thickness variation was assumed to be 4% of the nominal value and it was equally split amongst all three components of variation, namely, global variation, spatially correlated variation and independent variation. As the real measurement data for thickness correlation was unavailable, the covariance matrix for thickness variations used in this work was derived from an exponential decaying function of the respective distance. The correlation distance of exponential correlation function is normalized with respect to the chip dimensions.

Given the post-layout design implementation and a process variation model of oxide-thickness, the proposed methodology can compute the overall reliability distribution function. To validate the results of the proposed method, the overall reliability distribution was also computed from 1000 samples of Monte Carlo ($abbrev.$ MC) simulations using the same oxide reliability model and thickness variation model. In Table I, a comparison of lifetime estimation for 1-fault-per-million parts and 10-faults-per-million parts between the proposed approach and Monte Carlo simulations is shown for 5 design circuits. The size of the circuit under test in terms of number of devices is given in the second column. To verify the robustness of the proposed approach with respect to correlation distance we tested our approach for three different values of correlation distance normalized with respect to the chip size (ρ_{dist}=0.25, 0.5, 0.75). As can be seen from columns 3-6, the proposed statistical method is in good agreement with the Monte Carlo analysis, with errors of 1$\sim$3% for 10-faults-per-million and 3$\sim$6% for 1-fault-per-million.

TABLE III

ACCURACY COMPARISON BETWEEN PROPOSED APPROACH AND MC SIMULATIONS FOR DIFFERENT GRID RESOLUTION FOR DESIGN B

B	Lifetime Estimation Error w.r.t. MC simulations					
(80K)	Rho-distance = 0.25		Rho-distance = 0.5		Rho-distance = 0.75	
Grid Size	1/million	10/Million	1/million	10/Million	1/million	10/Million
5x5	5.09%	3.77%	4.63%	4.23%	4.37%	2.17%
10x10	5.01%	3.67%	5.37%	4.06%	4.26%	2.84%
20x20	5.26%	3.58%	5.07%	3.25%	4.93%	2.28%
25x25	5.20%	3.23%	5.06%	2.93%	3.92%	2.06%

TABLE II

RUN TIME COMPARISON BETWEEN THE PROPOSED STATISTICAL APPROACH AND MC SIMULATIONS

circuit	Time (sec.) per Sample			Overall Time (sec.)		
	Sta.	MC / speed-up		Sta.	MC / speed-up	
A	0.016	1.34	84X	1.62	1338	826X
B	0.016	2.17	136X	1.59	2171	1365X
C	0.015	2.71	181X	1.50	2712	1808X
D	0.018	7.52	418X	1.80	7517	4176X
E	0.018	18.63	1035X	1.78	18625	10463X

Table II presents the run time of the proposed statistical approach[1] ($abbrev$. Sta. in the table) as well as the Monte Carlo method. Unlike MC simulations, the proposed approach is able to analyze all the circuits in seconds. Columns 2-4 compares the run time per sample chip. The proposed approach demonstrates around 2∼3 orders of speed-up for different designs, whereas MC method scales super-linearly with the number of devices. Columns 5-7 document the overall run time for the ensemble of samples. The proposed approach shows four orders of improvement for the last design E. Moreover, as we discussed in Section 4.3, columns 2 and 5 confirm that the run time of the proposed approach is independent of the number of devices, which is an appealing feature as we face increasingly large circuit designs.

We also validates the approach by choosing 4 different resolutions of grid size for design B. The numerical results found for 4 different gird-size are given in Table III. As the discretization error of the grid-based model decreases for larger grid size, it can be seen that the error in estimation of reliability function also decreases in general.

Figure 8 shows the plot of overall $R_c(t)$ estimation computed using (1) Proposed approach (2) MC Simulation (3) Minimum oxide-thickness (Pessimistic approach) (4) Maximum oxide-thickness (Optimistic approach) for design C. It can be seen that reliability distribution function result for the proposed approach is in good agreement with MC result whereas the curves corresponding to the pessimistic and optimistic approaches deviate significantly from the proposed approach. It is noted that the proposed approach has relative errors of 3% and 2.8% for 10-faults-per-million and 50-faults-per-million. On the contrary, the approaches with minimum and maximum oxide-thickness lead to larger errors, 34.3% and 59.7% for 10-faults-per-million and 37.8% and 61.3% for 50-faults-per-million. This clearly exemplifies the need for a statistical approach for reliability distribution function analysis.

For examining the gaussian-assumption made for the oxide-thickness distribution of a chip, we computed the oxide-thickness distribution for the sample using MC simulations. Figure 9 shows the plots of oxide-thickness distributions for the samples of design C for different values of normalized correlation distance. It is clear that the gaussian distribution is a good fit for the oxide-thickness

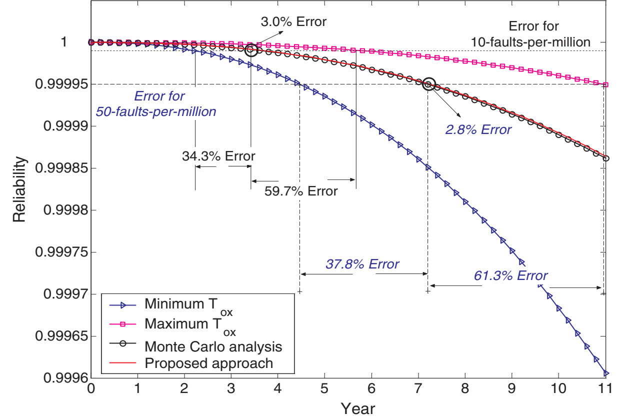

Fig. 8. Errors of the 10-faults-per-million and the 50-faults-per-million for the proposed approach, Monte Carlo simulations, worst-case oxide-thickness and best-case oxide-thickness

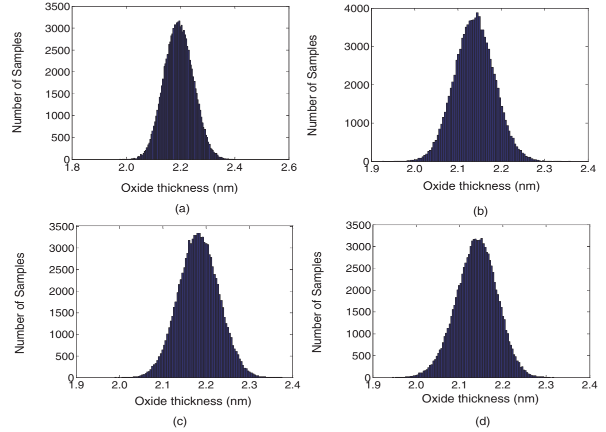

Fig. 9. Oxide-thickness distribution generated by MC simulations for Design C (100K gates, Grid size 20 × 20) for (a) $\rho_{dist} = 0.25$ (b) $\rho_{dist} = 0.5$ (c) $\rho_{dist} = 0.75$ (d) $\rho_{dist} = 1$

distribution. Likewise, we also varied the ratio of each component of thickness variation, the grid size and the chip size and similar results were observed.

Finally, we compare the difference between performing reliability analysis with and without considering actual spatial correlation between device oxide thickness. A set of MC simulations were done assuming accurate spatial correlation, no spatial-correlation and assuming perfect spatial correlation for gate-oxide thickness. Figure 10 compares the results with different spatial correlations for gate-oxide thickness. It can be observed that the mean values of the uncorrelated case is much smaller than the correlated cases (MC) and the perfectly correlated case overestimates the variance of the reliability function.

[1] Since the difference in correlation distance has no impact on the run time, Table II just cites the run time of the experiments with a correlation distance of 0.75.

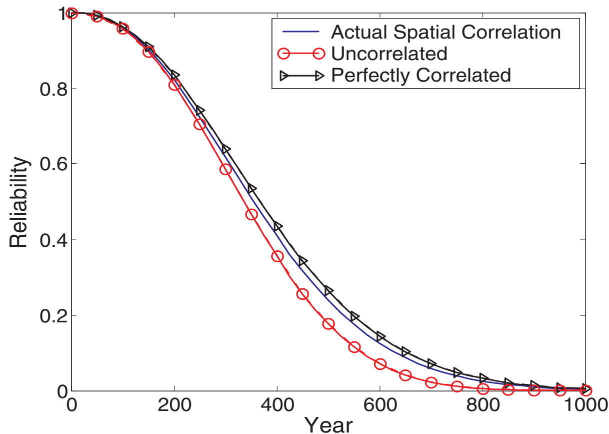

Fig. 10. Comparison between MC simulations of reliability function estimation with actual spatial correlation, perfect correlation and zero correlation

Therefore, statistical oxide reliability analysis without considering correlation may incorrectly predict the real lifetime of the circuit and could even underestimate the lifetime of the circuit.

VI. Conclusion

In this article, a statistical methodology for performing full chip oxide reliability analysis has been proposed, considering all three components of oxide-thickness variation. It is shown that worst-case oxide reliability analysis may not be adequate to predict chip lifetime accurately. The complexity analysis of the proposed methodology shows that the proposed approach is independent of the number of devices and is thus scalable to large industrial size circuits. Our simulation results exemplifies the accuracy and efficiency of the proposed method.

Acknowledgment

The authors gratefully acknowledge the Gigascale Systems Research Center (GSRC), Semiconductor Technology Academic Research Center (STARC) and National Science Foundation (NSF) for supporting this work.

References

[1] C. Hu. Gate oxide scaling limits and projection. In *Proc. IEDM*, pages 319–322, 1996.

[2] B. Kaczer, F. Crupi, R. Degraeve, P. Roussel, C. Ciofi, and G. Groeseneker. Observation of hot-carrier-induced nFET gate-oxide breakdown in dynamically stressed CMOS circuits. In *Proc. IEDM*, pages 171–174, 2002.

[3] Y. 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*, pages 18–28, 2006.

[4] J. Sune. New physics-based analytic approach to the thin-oxide breakdown statistics. *IEEE Electron Device Letter*, 22:296–298, 2001.

[5] H. Chang and S. Sapatnekar. Statistical timing analysis considering spatial correlations using a single pert-like traversal. In *Proc. ICCAD*, pages 621–625, 2003.

[6] A. Agarwal, D. Blaauw, V. Zolotov, S. Sundareswaran, M. Zhou, K. Gala, and R. Panda. Statistical delay computation considering spatial correlations. In *Proc. ASP-DAC*, pages 271–276, 2003.

[7] J. Xiong, V. Zolotov, and L. He. Robust extraction of spatial correlation. In *Proc. ISPD*, pages 2–9, 2006.

[8] A. Agarwal, D. Blaauw, and V. Zolotov. Statistical timing analysis for intra-Die process variations with spatial correlations. In *Proc. ICCAD*, pages 900–907, 2003.

[9] C. Visweswariah, K. Ravindran, K. Kalafala, S. Walker, and S. Narayan. First-order incremental block-based statistical timing analysis. In *Proc. DAC*, pages 331–336, 2004.

[10] Y. Lee, R. Nachman, S. Hu, N. Mielke, and J. Liu. Implant damage and gate-oxide-edge effects on product reliability. In *Proc. IEDM*, pages 481–484, 2004.

[11] E. Avni and J. Shappir. A model for silicon-oxide breakdown under high field and current stress. *Journal of Applied Physics*, 64:734–742, Jun. 1988.

[12] J. Stathis. Physical and predictive models of ultra thin oxide reliability in cmos devices and circuits. *IEEE Trans. on Devices and Materials Reliability*, 1:43–59, 2001.

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

[14] J. Sune and E. Y.Wu. Statistics of successive breakdown events in gate oxides. *IEEE Electron Device Letter*, 24(Issue 4):272–274, 2003.

[15] J. P. Imhof. Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48:419–426, 1961.

[16] K.-H. Yuan and P. M. Bentler. Two simple approximations to the distributions of quadratic forms. *Department of Statistics, UCLA. Department of Statistics Papers.* Paper 2007010106. Available at: http://repositories.cdlib.org/uclastat/papers/2007010106.