Process Variation and Temperature-Aware Full Chip Oxide Breakdown Reliability Analysis

Cheng Zhuo, *Student Member, IEEE,* Kaviraj Chopra, Dennis Sylvester, *Fellow, IEEE,* and David Blaauw, *Senior Member, IEEE*

***Abstract*—Gate oxide breakdown (OBD) is a key factor limiting the useful lifetime of an integrated circuit. Unfortunately, the conventional approach for full chip OBD reliability analysis assumes a uniform oxide thickness and worst-case temperature for all devices. In practice, however, gate oxide thickness varies from die-to-die and within-die and hence may cause different reliability for different devices even chips. Moreover, due to the increased across-die temperature variation, such difference may be exacerbated. Thus, as the precision of variation control worsens, an alternative reliability analysis approach is needed. In this paper, we propose a statistical framework for chip-level gate OBD reliability analysis while considering both die-to-die and within-die components of thickness variations as well as the across-die temperature 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 chip-level OBD reliability function 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 much fewer 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 as well as temperature variation. We show that, compared to Monte Carlo simulation, the proposed method incurs an error of only around 1% while improving the runtime by more than three orders of magnitude.**

***Index Terms*—Oxide breakdown, process variation, reliability, spatial correlation, temperature.**

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 [1]. 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, and others, 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 few decades, numerous publications have focused on understanding and modeling the mechanisms leading to defect generation and breakdown in individual devices [2], [3]. Some researchers have initiated an effort to understand the oxide breakdown (OBD) mechanisms of simple circuits [3]. Recently, a product level approach performing OBD analysis on full chip was proposed in [4]. In most of the existing approaches, simple test structures such as discrete devices or capacitors are used to characterize the OBD 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, there are two major concerns associated with the prior approaches.

1) Prior approaches assume a *uniform minimum* 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 [5], [6]. Therefore, in previous approaches, it was imperative to consider a uniform minimum oxide thickness across all devices on a chip and across all chips for a conservative worst-case analysis. This may lead to significantly pessimistic estimates of the overall OBD reliability of the product.
2) In addition, prior works assume a *worst* operating temperature across the chip and throughout the lifetime. However, it has been well noted that the temperature varies significantly across the chip. Since devices operated at different temperature may deteriorate at different rates, the mean time to failure of a device actually exponentially depends on temperature [7]–[9]. Fig. 1(a) and (b) illustrates the temperature profiles for a traditional alpha processor [10] and a modern many-core design

Manuscript received June 25, 2010; revised November 23, 2010 and February 5, 2011; accepted March 27, 2011. Date of current version August 19, 2011. This work was supported by the National Science Foundation. This paper was recommended by Associate Editor D. Z. Pan.

C. Zhuo, D. Sylvester, and D. Blaauw are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: czhuo@umich.edu; dennis@eecs.umich.edu; blaauw@umich.edu).

K. Chopra is with Mentor Graphics, Inc., San Jose, CA 95131 USA (e-mail: kaviraj_chopra@mentor.com).

Digital Object Identifier 10.1109/TCAD.2011.2142183

1089-778X/$26.00 © 2011 IEEE

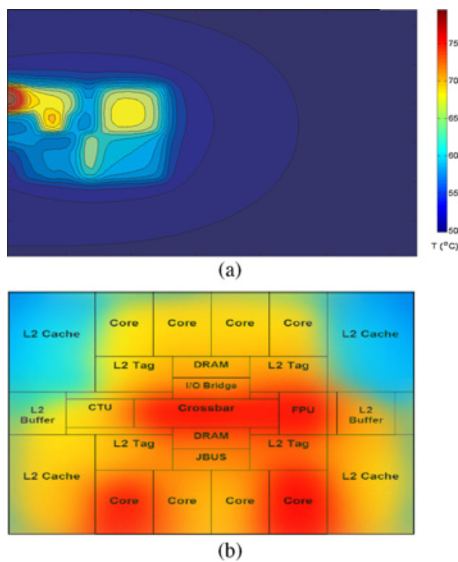

Fig. 1. (a) Temperature profile for an alpha processor by HotSpot [10]. (b) Temperature profile for a many-core processor from [11] and [12].

from [11] and [12]. It is easy to observe that the hot spots only occupy a small region of the entire chip and have around 30 degrees of temperature difference from the inactive regions. For a device in inactive regions, such temperature difference may lead to the reliability underestimation by one order of magnitude [7], [8], which is overly pessimistic.

Since 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 limits, it is critical to quantify the product OBD strength with consideration of both process and temperature variations.

The goal of this paper is to develop a new chip-level gate OBD reliability analysis while accommodating both *process* (inter-die, intra-die spatially correlated and independent) and *temperature variations* into the analysis flow. 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 multidimensional nested stochastic process. Furthermore, temperature variations in hot spots and inactive regions may result in different device-level reliability models, which also complicates the reliability analysis.

Apparently a straightforward Monte Carlo (MC) approach is extremely expensive in both execution time and memory, as we need to perform nested MC analysis on the sample spaces for different chips and different devices across each chip as well as the sample space of OBD of each device. Meanwhile, the traditional guard-band approach by assuming the minimum oxide thickness may reach overly pessimistic result with more than 50% underestimation [14]. The challenge here is how to reduce the tremendous number of random variables for a low space/time complexity while maintaining good reliability analysis accuracy. The contributions of the proposed framework are as follows.

1) First, we present a more consistent and accurate model for statistical full-chip reliability analysis. Unlike any traditional reliability analysis that simply uses the worst corner, the proposed model incorporates both the oxide thickness variation and temperature variation to ensure a reasonable result. This theoretically rigorous OBD reliability model considers variations at different spatial scales and hence involves the integration over millions of variables, which may be difficult for a direct solve.
2) Second, the proposed framework discusses how to project that tremendous parameter space at device-level to the granularity of block level by characterizing the *block-level oxide thickness distribution* (BLOD). Fig. 1(a) and (b) illustrates the global temperature unevenness (corresponding to different functional modules) and local temperature uniformity for two processors [10]–[12]. Based on such observation, a "block" is defined as a region on chip with uniform temperature spread.[1] We therefore present how to map the millions of random variables within each block to only two distinct random variables of sample mean and sample variance of BLOD. Such a projection greatly reduces the problem size to a feasible level while still capturing both the temperature variation and the process variations at different spatial scales.
3) Third, we demonstrate how to characterize the sample mean/variance for each block using principal components and then compute the full-chip OBD reliability in an efficient way. By expressing the oxide thickness variation with principal components, we can achieve the closed-form representation of the sample mean/variance for each block as well as their correlation. Then, with some judicious approximations, the initial high-dimensional integration for the reliability across the ensemble of chips can be simplified to the sum of double integrals, which enables fast and accurate estimation.

The remainder of this paper is organized as follows. In Section II, we describe the oxide thickness variation modeling at different spatial scales. In Section III, we discuss the OBD model and formulate the oxide reliability analysis problem. Section IV details the proposed variation-aware full chip OBD reliability analysis. Simulation results illustrating the efficacy of the proposed approach are given in Section V. We conclude this paper in Section VI. A part of this paper has been published in [16] and [17].

II. Review on Oxide Thickness Variation Modeling

The oxide thickness variation can be classified based on the spatial scale over which it manifests [18]–[20]. Due to long

[1]This "block" could be a real architecture level block or some sub-block that can ensure the assumption of uniform temperature. In general, tens of blocks can capture the major feature of the thermal profile [13].

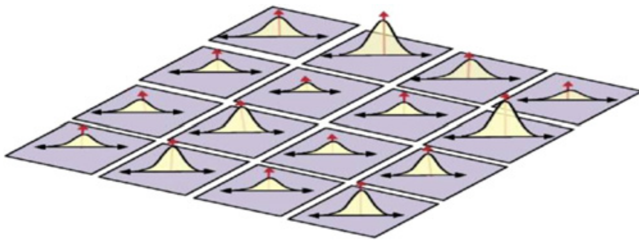

Fig. 2. Grid-based spatial correlation model.

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 die-to-die component of variations is referred to as *global* or *inter-die 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*. Thus, oxide thickness for any device can be modeled as follows:

$$x = u_0 + z_g + z_{corr} + z_\epsilon \tag{1}$$

where u_0 is the nominal oxide thickness for the technology, z_g denotes the global inter-die variation component, z_{corr} is the intra-die spatially correlated component that tends to affect closely placed devices in a similar manner, and z_ϵ is the residual 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 [20]. Therefore, simplified correlation structures using a grid model [18], [20] or quad-tree model [24] have been proposed in the literature. In this paper, 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 Fig. 2), and a covariance matrix of size $n \times n$ representing the spatial correlations among the grids.[2] The covariance matrix could be determined from measurement data extracted from manufactured wafers using the method given in [20]. Some recent researches [21]–[23] notice that part of the intra-die spatially correlated variation could be attributed to the wafer-level global deterministic pattern (e.g., slanted or bowl-shaped). Such pattern is usually characterized by a quadratic or some polynomial functions [21], [23]. Given the locations of the chip and the grids within the chip, the models by [21] and (1) could be compatible by replacing the common inter-die variation component z_g in (1) with a location-dependent component for each grid.

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 [18], [19], where oxide thickness x of any device in ith grid is given by

$$x = \lambda_{i,0} + \sum_{j=1}^{n} \lambda_{i,j} z_j + \lambda_r \epsilon \tag{2}$$

where $\lambda_{i,0}$ is the mean or nominal value of oxide thickness in ith 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 ith principal component for every jth the random variable.

III. Reliability Model and Problem Formulation

The gate oxide degradation depends on the oxide thickness, voltage, and temperature. There are many OBD models in the literature that attempt to explain the dependence on these factors. A widely accepted model is the anode hole injection model [25]. 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 OBD [26]. Both models of OBD mechanisms note that the defect generation is a non-deterministic process. As a result, the OBD time is inherently a statistically distributed quantity. Thus, the OBD time is modeled as a random variable typically characterized by a Weibull probability distribution function, given by [5] and [27] as follows:

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

where F is the cumulative distribution function (CDF) of time-to-breakdown t, a is the device area normalized with respect to (w.r.t.) the minimum device area, and 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

[2]The "grid" in the correlation model may be different from the "block" that is defined earlier and used for temperature uniformity partition. Thus, a block with uniform temperature spread may contains several grids for a finely gridded model or be contained within one grid for a coarsely gridded model.

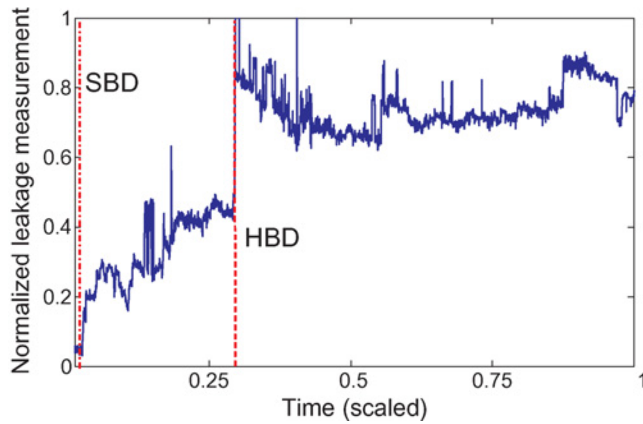

Fig. 3. Typical OBD procedure characterized by gate leakage for a device in 45 nm process (the stressed condition is 3.1 V, 100 °C).

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 [6]. Thus, if x denotes the oxide thickness, we have

$$F(t) = 1 - e^{-a(\frac{t}{\alpha})^{bx}} \tag{4}$$

where b is a constant for given temperature and supply voltage. It has also been noted that the parameters α and b depend on temperature and can be characterized using some closed-form models or look-up tables w.r.t. temperature for a given process [7]–[9]. Calibration and measurement results in [4], [6]–[9], and [28] validate the underlying Weibull model for OBD mechanism and also demonstrate its impact on designs with different dimensions. Another major factor that affects the oxide lifetime is the OBD 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 (HBD) [28]. The time between two breakdowns is a function of the gate area, oxide quality, and the bias conditions [4], [28]. Fig. 3 plots the gate leakage measurement results for a stressed device in 45 nm process. It can be observed from the figure how the gate leakage continuously increases after SBD until HBD is triggered. Thus, SBD is an irreversible process and may change the gate leakage by 10–20 times, or even several orders for some central processing unit (CPU) design [4]. Such significant leakage increase may easily lead to cache failure, which dominates the CPU lifetest fallout. It is also worth noting that the leakage increase does not necessarily result in circuit/logic failure and circuit may even survive to function after several HBDs [4], [29], [30]. In other words, the selection of the failure criterions may also depend on the application and the design under investigation. For the purpose of this paper to enable the reliability analysis for large chips especially CPU designs, we limit our analysis to determining the initiation of SBD and use this as our failure criteria.

To ensure the robustness, 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

$$R(t) = 1 - F(t) = P(T > t) = \int_t^\infty f(s)ds \tag{5}$$

where $f(s)$ is the probability density function (PDF) of OBD of an individual device. The reliability function is complimentary to the 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

$$R_i(t|x_i) = P(t > \mathbf{t}|x_i) = \int_t^\infty f(s|x_i)ds. \tag{6}$$

Due to the spatial correlation 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 *a priori* then the defect generation mechanism in one device is independent of any other device on the chip for constant 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 conditional reliability function of the chip $R_c(t|\mathbf{x})$ (with oxide thicknesses known) requires that all devices on the chip are functioning reliably. Thus, $R_c(t|\mathbf{x})$ is given by the product of reliability functions of all individual devices as follows:

$$R_c(t|\mathbf{x}) = \prod_{i=1}^{m} R_i(t|x_i) \tag{7}$$

where $\mathbf{x}$ represent the vector of oxide thickness $(x_1, \dots, x_m)$ 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 (7) 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. In addition, these random variables are correlated across the chip. Furthermore, due to the temperature variation, many devices may have different reliability functions. If the oxide thicknesses of all devices are characterized by their joint PDF $f(x_1, \dots, x_m)$ then the overall reliability function of the entire ensemble of all manufactured chips can be given by

$$R_c(t) = \int_0^\infty \dots \int_0^\infty \prod_{i=1}^{m} R_i(t|x_i) \qquad f(x_1, \dots, x_m) \; dx_1, \dots, dx_m. \tag{8}$$

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 certain projection techniques, we develop a computationally efficient approach to address this problem in the next section.

IV. Process Variation and Temperature-Aware Full Chip OBD 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 one device. Using this expression, the conditional reliability function of a particular chip can be found given the oxide thickness of all devices on the chip as well as the temperature profile. 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 first partition the chip into the granularity of N blocks, in which devices share similar temperatures. Then, within each block, 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[3] and there are a large number of devices within a block, we show that such frequency distribution across a given block can be approximated by a normal distribution function [31]. Henceforth, we will refer to this distribution function as the BLOD. The BLOD allows us to compactly represent the oxide thickness of all device within a block of a given chip using just two parameters: the mean and the variance of the underlying normal distribution function.

In Section IV-A, we will present how a closed-form expression for the block-level reliability function can be found for one BLOD. Then the chip-level reliability for a given chip can also be analytically computed from N BLODs. Finally, we discuss how to compute the overall reliability function across the entire ensemble of all manufactured chips. As N BLODs may vary from die to die, their means and variances are in fact random variables over the sample space of all manufactured chips. Hence, the means and variances of N BLODs can be represented by two random vectors, with N entries corresponding to N BLODs in each vector. In other words, several million multivariate oxide thickness distribution function for each device on the chip can be compactly modeled with just two random vectors. In Section IV-C, we will discuss how these two random vectors can be derived from the oxide thickness process variation model given in (2) and thus the overall reliability function can be computed from it.

For clarity, we define the following notations in Table I that will be used throughout the remainder of this paper.

[3]The normal random variable model for oxide thickness variation could be validated by solving the general equations for the rate of growth of the oxide, as in [31].

TABLE I

Notations Used in OBD Reliability Analysis

Notation	Definition
N	No. of functional blocks in a chip
m	No. of devices of a chip
m_j	No. of devices in the j_{th} block, i.e., $\sum_{j=1}^{N} m_j = m$
n	No. of grids in the spatial correlation model of (2)
$\mathbf{x} = [x_1, \ldots, x_m]$	The oxide thicknesses for m device of a chip
$x_{i,j}$	Oxide thickness for the i_{th} device in the j_{th} block of a chip, $i = 1, \ldots, m$
$a_{i,j}$	Area for the i_{th} device in the j_{th} block of a chip, $i = 1, \ldots, m$
A_j	Total area for the j_{th} block of a chip, $j = 1, \ldots, N$
$\bar{x}_j = \sum_{i=1}^{m_j} x_{i,j}/m_j$	The sample mean for m_j devices of the j_{th} block
$v_j = \frac{\sum_{i=1}^{m_j}(x_{i,j}-\bar{x}_j)^2}{m_j-1}$	The sample variance for m_j device of the j_{th} block
$f_{\mathbf{x},\mathbf{y}}(x, y)$	Joint PDF of x and y, where x and y can be either vector or scalar

A. OBD Reliability Analysis for One Chip

Using the definition of the reliability function and the OBD time model of an individual device in (4), the conditional reliability function of an ith device on a chip having oxide thickness x_i is given by

$$R_i(t|x_i) = e^{-a_i(\frac{t}{\alpha_i})^{b_i x_i}} \tag{9}$$

where α_i and b_i are temp device-level reliability parameters for the i_{th} device.

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 $\mathbf{x} = [x_1, x_2, \ldots, x_m]$ and their respective area a_i, the conditional reliability of the chip is given by

$$R_c(t|\mathbf{x}) = \prod_{i=1}^{m} R_i(t|x_i) = e^{-\sum_{i=1}^{m} a_i(\frac{t}{\alpha_i})^{b_i x_i}}. \tag{10}$$

There may be several million devices on a chip and parameters of α_i and b_i may differ. Thus, 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 while considering the impact of temperature variation across the chip.

It has been noted in Fig. 1(a) and (b) that the on-chip temperature profile has the characteristics of global difference and local uniformity. Since both parameters b and α are heavily dependent on temperature [7]–[9], it is therefore unfair to assume that hot spots and inactive areas have the same reliability model and are hence equally prone to the OBD failure. In practice, temperature profile of a chip varies continuously across the chip. Transistors within a particular block may share similar temperature due to the similar activities and supply voltage [10], [15]. On the contrary, temperature variation from block to block is much higher as functional blocks usually perform completely different operations [15]. It is therefore sufficient to construct a temperature-aware reliability

analysis model at the granularity of *blocks*, within which the temperature spread is uniform. In other words, the analysis reasonably depends on the fact that devices within a block may bear similar temperature and hence share approximately the same parameters α_i and b_i for the reliability functions. As a result, we considers the block-level worst-case operating temperature and supply voltage in the analysis to account for the block-level temperature difference and to ensure a correct operation throughout the entire life time for any application profile. Then, (10) can be expressed at the functional block level as follows:

$$R_c(t|\mathbf{x}) = \prod_{i=1}^{m} R_i(t|x_i) = e^{-\sum_{j=1}^{N}\sum_{i=1}^{m_j} a_{i,j}(\frac{t}{\alpha_j})^{b_j x_{i,j}}} \tag{11}$$

where N is the number of architecture-level blocks, and α_j and b_j denote the parameters of the reliability functions for devices in the j_{th} block.

Equation (11) considers the across-chip temperature variation but cannot simplify the model. To achieve that, we represent the set of devices within a block and their individual oxide thicknesses using BLOD for a particular block in a chip. For example, for the j_{th} block, the shape of its BLOD can be approximated by performing histogram of the oxide thicknesses of $x_{i,j}$ for all the devices within that block. This block-specific BLOD shows how many devices correspond to a particular oxide thickness within that block. For the sake of understanding, we discretize this oxide thickness distribution for the j_{th} block into k_j discrete intervals assuming a truncated distribution. It can be seen that when we make this transformation the area of the devices with identical thickness in a block can be summed together. Let $\overline{x}_{i,j}$ denote the midpoint of the ith discrete interval for the j_{th} BLOD and $\overline{a}_{i,j}$ be the total area of all devices having thickness $\overline{x}_{i,j}$ in that block. By applying this transformation, the above expression for $R_c(t|\mathbf{x})$ can be rewritten as

$$R_c(t|\mathbf{x}) = e^{-\sum_{j=1}^{N}\sum_{i=1}^{k_j} \overline{a}_{i,j}(\frac{t}{\alpha_j})^{b_j \overline{x}_{i,j}}}. \tag{12}$$

By making such a transformation, the dimensionality of $R_c(t|\mathbf{x})$ can be reduced from number of devices m to the sum of the number of discrete intervals k_j, i.e., $\sum_{j=1}^{N} k_j$. If we normalize the exponent with total area of each block, the above expression gives

$$R_c(t|\mathbf{x}) = \exp\left[-\sum_{j=1}^{N} A_j \sum_{i=1}^{k_j} p_{i,j}(\frac{t}{\alpha_j})^{b_j \overline{x}_{i,j}}\right] \tag{13}$$

where $p_{i,j} = \overline{a}_{i,j}/A_j$ represents the probability of observing an oxide thickness $x_{i,j}$ on a particular block of a sample chip. Thus, the thickness of all devices on a particular sample chip can be compactly characterized by N BLODs.

As discussed in Section II, the thickness variation of a device includes global variation (inter-die), spatially correlated intra-die variation (modeled as multivariate Gaussian random vector for devices across the chip), and random variation [residual component, modeled as an independent Gaussian random variable, e.g., $N(0, \sigma_\epsilon^2)$] [18], [31]. Thus, for a set of devices within one particular block, they may have different oxide thicknesses due to the variability. The BLOD is interpreted as the frequency distribution histogram of observing certain oxide thicknesses in this block for a sample chip, which has the following property.

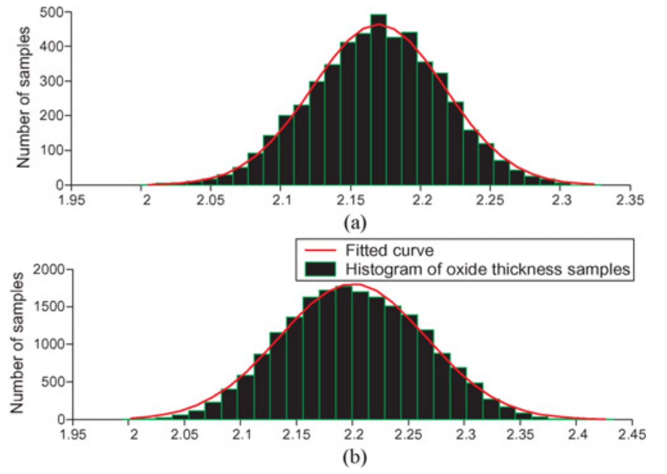

Fig. 4. (a) Histogram of the oxide thickness for a block with 5K devices. (b) Histogram of the oxide thickness for a block with 20K devices.

Property: Following the oxide thickness variation classification above and the discussion in [31], BLOD can be approximated by the curve of a Gaussian distribution.

This property could be understood by analyzing the impact of different variation components as follows.

1) First, within one block, all the devices share the same global variation component (z_g).
2) Second, devices within the same block are closely placed and hence highly or even perfectly spatially correlated. This implies they may have approximately the same spatially correlated variation component ($z_{corr,j}$).
3) Third, by using the variation component classification of inter-die, spatially correlated intra-die and random variations in (1), the oxide thickness of a device in the j_{th} block is

$$x_{i,j} = u_0 + z_g + z_{corr,j} + z_{\epsilon,i,j}. \tag{14}$$

For any device in the block, u_0, z_g, and $z_{corr,j}$ are approximately the same. The difference of oxide thicknesses is therefore mainly caused by the random variation component $z_{\epsilon,i,j}$. In other words, oxide thickness of any device within the block can be considered as a sample from a Gaussian process $N(u_0 + z_g + z_{corr,j}, \sigma_\epsilon^2)$. Due to the independence of the random variation component $z_{\epsilon,i,j}$, oxide thicknesses of devices within one block are simply samples independently drawn from one common random process $N(u_0 + z_g + z_{corr_j}, \sigma_\epsilon^2)$.

As long as the number of devices is sufficient, BLOD can be well characterized by the histogram of oxide thickness samples from the Gaussian random process $N(u_0{+}z_g{+}z_{corr,j}, \sigma_\epsilon^2)$, which hence follows the curve of a Gaussian distribution.[4]

[4]Even if for the particular case that this approximation does not work well, we still can include more moments and pick up an appropriate distribution to describe BLOD.

This property helps shed insight into the shape of a BLOD and how $p_{i,j}$ may change w.r.t. oxide thickness x. Fig. 4 validates the property with the histograms of oxide thicknesses for two blocks with different number of devices by MC simulation. It is clear that either for a block with 5K devices [Fig. 4(a)] or a block with 20K devices [Fig. 4(b)], we get distinctly Gaussian-like curves with very high fitting goodness (R-square) of 99.8% and 99.5%, respectively, which qualitatively justify the property. Thus, the summation part $\sum_{i=1}^{k_j} p_{i,j}(\frac{t}{\alpha_j})^{b_j\bar{x}_{i,j}}$ in (13) can be expressed by the integration over x, as the number of devices within each block is usually sufficient to ensure the convergence as follows:

$$R_c(t|\mathbf{x}) = R_c(t|\mathbf{u},\mathbf{v}) \approx \prod_{j=1}^{N} \exp[-A_j \int_{-\infty}^{\infty} \phi(\frac{x-u_j}{\sqrt{v_j}})(\frac{t}{\alpha_j})^{b_j x} dx] \quad (15)$$

where $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is the PDF for a standard Gaussian distribution. $\mathbf{u} = (u_1, u_2, \ldots, u_N)$ and $\mathbf{v} = (v_1, v_2, \ldots, v_N)$, where u_j and v_j are the sample mean and variance of the j_{th} BLOD.

Since (13) computes the conditional reliability $R_c(t|\mathbf{x})$ using only $2N$ distinct variables, the dimensionality of the problem in (10) is reduced from millions to $2N$. However, microprocessors usually have tens of blocks, making an integration with $2N$ variables still difficult to solve for (8). Note that $\exp[-A_j \int_{-\infty}^{\infty} \phi(\frac{x-u_j}{\sqrt{v_j}})(\frac{t}{\alpha_j})^{b_j x} dx]$ is approximately the product of all the device-level reliability functions in the j_{th} block, and hence very close to 1 within the lifetime of interest. By applying first-order Taylor expansion, (13) can be further simplified to

$$\begin{aligned} &R_c(t|\mathbf{u},\mathbf{v}) \\ &= \prod_{j=1}^{N} 1 - \left[1 - e^{-A_j \int_{-\infty}^{\infty} \phi(\frac{x-u_j}{\sqrt{v_j}})(\frac{t}{\alpha_j})^{b_j x} dx}\right] \\ &\approx 1 - \sum_{j=1}^{N} \left[1 - e^{-A_j \int_{-\infty}^{\infty} \phi(\frac{x-u_j}{\sqrt{v_j}})(\frac{t}{\alpha_j})^{b_j x} dx}\right]. \end{aligned} \quad (16)$$

In the above equation, the integral in the exponent can be analytically evaluated by making the substitution $\frac{t}{\alpha_j} = e^{\gamma}$ as follows:

$$\begin{aligned} &\int_{-\infty}^{\infty} \phi(\frac{x-u_j}{\sqrt{v_j}})(\frac{t}{\alpha})^{b_j x} dx = \int_{-\infty}^{\infty} \phi(\frac{x-u_j}{\sqrt{v_j}}) e^{\gamma b_j x} dx \\ &= -\frac{1}{2} e^{\gamma b_j u_j + \gamma^2 b_j^2 v_j/2} erf(\frac{-x+u_j+\gamma b_j v_j}{\sqrt{2 v_j}})|_{-\infty}^{\infty} \\ &= e^{\ln(\frac{t}{\alpha_j}) b_j u_j + (\ln(\frac{t}{\alpha_j}))^2 b_j^2 v_j/2}. \end{aligned} \quad (17)$$

Equation (17) is denoted as $g(u_j, v_j)$ for simplicity throughout the remainder of this paper.

Thus, for N given BLODs $\phi(\frac{x-u_j}{\sqrt{v_j}})$, where $j = 1, \ldots, N$, the conditional reliability function of a chip can be computed by the closed-form expression as follows:

$$\begin{aligned} &R_c(t|\mathbf{u},\mathbf{v}) \\ &= 1 - \sum_{j=1}^{N} \left[1 - e^{-A_j e^{\ln(\frac{t}{\alpha_j}) b_j u_j + (\ln(\frac{t}{\alpha_j}))^2 b_j^2 v_j/2}}\right]. \end{aligned} \quad (18)$$

Hence, the multidimensional exponent in (10) can now be compactly represented using a closed-form analytical function of BLOD parameters $\mathbf{u}$ and $\mathbf{v}$.

B. Design-Time OBD Reliability Analysis for the Ensemble of Chips

At design time when chips are not fabricated, designers are unable to know the specific BLOD distribution for any block of any chip. In other words, $\mathbf{u}$ and $\mathbf{v}$ turn out to be two random vectors at design time. The OBD reliability is then evaluated for the design (or the ensemble of chips) instead of a particular chip by integrating the conditional reliability function in (18) over the joint PDF of random vectors $\mathbf{u}$ and $\mathbf{v}$. In this section, we will discuss how to achieve a compact expression of the overall reliability function by enumerating the conditional reliability function derived in the previous section across the ensemble of all chips.

As shown in Fig. 5, each sample chip from one design may result in different BLODs for the same block from chip to chip, therefore, the oxide thickness variation of one block across the entire ensemble of all chips can be represented with a set of BLODs for all manufactured chips. Now each such BLOD is characterized by their respective mean u_j and variance v_j. Therefore, the oxide thickness distribution of all devices across all manufactured chips with N blocks can be represented by $2N$ random variables $\mathbf{u} = [u_1, u_2, \ldots, u_N]$ and $\mathbf{v} = [v_1, v_2, \ldots, v_N]$. In other words, for one particular chip, its BLODs simply result from the samples of two random vectors.

Now let $f_{\mathbf{uv}}(\mathbf{u},\mathbf{v})$ denote the joint PDF of $\mathbf{u}$ and $\mathbf{v}$. For computing the overall reliability function, we need to integrate the above expression of reliability function of one chip over the joint PDF $f_{\mathbf{uv}}(\mathbf{u},\mathbf{v})$ of random vectors $\mathbf{u}$ and $\mathbf{v}$ as follows:

$$\begin{aligned} R_c(t) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left[1 - \sum_{j=1}^{N} \left(1 - e^{-A_j g(u_j, v_j)}\right)\right] \\ &\times f_{\mathbf{u},\mathbf{v}}(\mathbf{u},\mathbf{v}) du_1 \ldots du_N dv_1 \ldots dv_N \\ &= 1 - N + \sum_{j=1}^{N} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-A_j g(u_j, v_j)} \\ &\times f_{\mathbf{u},\mathbf{v}}(\mathbf{u},\mathbf{v}) du_1 dv_1 \ldots du_N dv_N \end{aligned} \quad (19)$$

where $g(u_j, v_j)$ is defined in (17). Since $\exp[-A_j g(u_j, v_j)]$ is independent of any other u_i or v_i that $i \neq j$, we have

$$\begin{aligned} &\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-A_j g(u_j, v_j)} f_{\mathbf{u},\mathbf{v}}(\mathbf{u},\mathbf{v}) du_1 dv_1 \ldots du_N dv_N \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-A_j g(u_j, v_j)} \\ &\times \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{u},\mathbf{v}}(\mathbf{u},\mathbf{v}) du_1 dv_1 \ldots du_N dv_N \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-A_j g(u_j, v_j)} f_{\mathbf{u_j},\mathbf{v_j}}(u_j, v_j) du_j dv_j. \end{aligned} \quad (20)$$

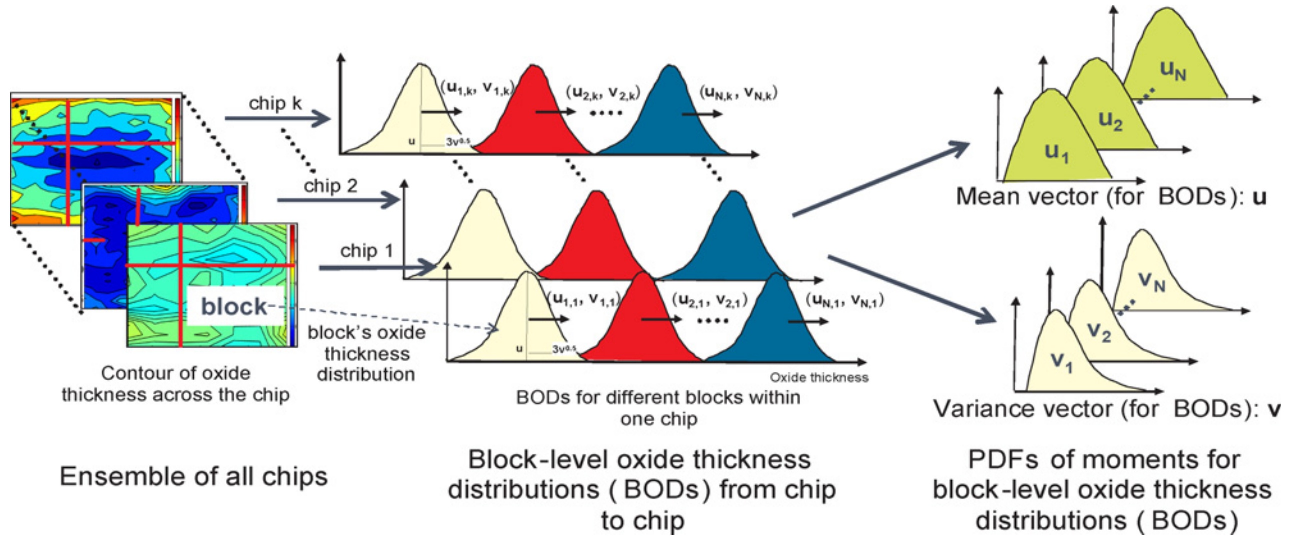

Fig. 5. Compact representation of oxide thickness variation for the ensemble of chips. (a) Ensemble of all chips. (b) BLODs from chip to chip. (c) PDFs of moments for BLODs.

Thus, we can express the design time OBD reliability for the ensemble of chips in (20) using N double integrals as follows:

$$R_c(t) = 1 - N + \sum_{j=1}^{N} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-A_j g(u_j, v_j)} f_{\mathbf{u_j}, \mathbf{v_j}}(u_j, v_j) du_j dv_j. \quad (21)$$

C. $f_{\mathbf{u_j}, \mathbf{v_j}}(u_j, v_j)$ Characterization Using Principal Component Analysis

To compute (21), it is still required to know the characteristics of joint PDF $f_{\mathbf{u_j}, \mathbf{v_j}}(u_j, v_j)$ of each BLOD. In this section, we will discuss how to characterize those joint PDFs for blocks by using principal component analysis (PCA).

As is discussed in the previous section, each BLOD approximately follows a Gaussian curve and hence can be represented by their respective mean u_j and variance v_j. For a particular block j, the mean u_j and variance v_j of its oxide thickness distribution can be estimated by calculating the unbiased statistical BLOD mean and variance of the oxide thickness values observed across the block. Likewise, the random variables u_j and v_j can be found in terms of the thickness variation model discussed in (2). Using the oxide thickness variation model given in (2), sample mean u_j can be expressed as follows:

$$u_j = \frac{1}{m_j} \sum_{i=1}^{m_j} x_{i,j} = u_{j,0} + \sum_{k=1}^{n} u_{j,k} z_k + u_{j,n+1} \epsilon \quad (22)$$

where m_j is the number of devices within each block and hence the number of "samples" to compute the sample mean and variance.

The grid-based model in (2) partitions the chip into several grids, as discussed in Section II. Assume the i_{th} device in the j_{th} block is located in a grid, e.g., grid $g_{i,j}$, where $g_{i,j}$ corresponds to a grid index from 1 to n. Then, we can compute $u_{j,k}$ and $u_{j,n+1}$ as follows:

$$u_{j,k} = \frac{1}{m_j} \sum_{i=1}^{m_j} \lambda_{g_{i,j},k} \quad \forall k = 0, \ldots, n$$

$$u_{j,n+1} = \frac{1}{m_j} \sqrt{\sum_{i=1}^{m_j} \lambda_r^2} = \frac{\lambda_r}{\sqrt{m_j}}.$$

The coefficient $u_{j,0}$ is the nominal value of u_j, whereas coefficient $u_{j,i}$ is the sensitivity to the i_{th} principal component. It is evident that the sensitivity of the independent random component $u_{j,n+1}$ tends to zero for a large number of devices and thus can be safely neglected for a typical industrial chip.

Similarly the expression for v_j, the sample variance of the j_{th} BLOD, in terms of oxide variation model in (2), can be given as follows:

$$v_j = \frac{1}{m_j - 1} \sum_{i=1}^{m_j} (x_{i,j} - u_j)^2 = \frac{1}{m_j - 1} \sum_{i=1}^{m_j} (x_{i,j}^2 - u_j{}^2). \quad (23)$$

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

$$v_j = v_{j,0} + \sum_{i=1}^{n} \sum_{k=1}^{n} v_{j,i,k} z_i z_k \quad (24)$$

where

$$v_{j,0} = \lambda_r^2 \quad \text{and} \quad v_{j,i,k} = \frac{1}{m_j - 1} \sum_{l=1}^{n} (\lambda_{l,i} - u_{j,i})(\lambda_{l,k} + u_{j,k}).$$

In this manner, we can express the distributions of u_j and v_j in terms of a given process variation model. Note that random variable u_j is the sum of normal random variables so it is also a normal random variable, however, the BLOD variance v_j is not a normal random variable. By exploring their characteristics, we have the following lemma for their un-correlation.

Lemma: Following the oxide thickness variation model in (2), u_j and v_j for a BLOD is uncorrelated, i.e., $E[u_j v_j] = E[u_j]E[v_j]$, where $E[\cdot]$ denotes the expectation.

Proof: Following the principal component models discussed above, we can express $E[u_jv_j]$ as follows:

$$\begin{aligned} E[u_jv_j] &= E[(u_{j,0} + \sum_{i=1}^{n} u_{j,i}z_i + u_{j,p+1}\epsilon) \\ &\quad \times(v_{j,0} + \sum_{i=1}^{n}\sum_{k=1}^{n} v_{j,i,k}z_iz_k)]. \end{aligned}$$

By noting that each principal component z_i as an independent standard normal random variable, we have

$$\begin{aligned} E[z_i] = E[z_i^2z_j] &= E[z_iz_j^2] = E[z_i^3] = 0 \\ E[z_i^2] &= 1 \end{aligned} \tag{25}$$

for different i and j. Likewise

$$\begin{aligned} E[\epsilon] = E[\epsilon^2z_j] &= E[z_i\epsilon^2] = E[\epsilon^3] = 0 \\ E[\epsilon^2] &= 1. \end{aligned} \tag{26}$$

Thus, the above expression can be simplified and given by

$$E[u_jv_j] = u_{j,0}v_{j,0} + \sum_{i=1}^{n} u_{j,0}v_{j,i,i} = E[u_j]E[v_j]. \tag{27}$$

■

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 v_j is not a normal random variable and has the distribution of quadratic forms in normal variables [32], [33]. However, we still can achieve the joint PDF for u_j and v_j in a numerical way by generating the MC samples of the principal components and then constructing the joint PDF with (22) and (24). As is observed in Fig. 1(a) and (b), it is usually sufficient to partition the design into tens of blocks (N) to accurately capture the temperature profile. Thus, even by constructing the joint PDF in a numerical way, the complexity is already significantly lower in comparison to (8).

Moreover, with numerical experiments we find that the dependence between u_j and v_j is weak. As a result, it is reasonable to assume u_j and v_j as independent variables, which allows us to express the joint PDF in terms of their marginal distributions $f_{\mathbf{u_j}}(u_j)$ and $f_{\mathbf{v_j}}(v_j)$ and further reduce the complexity. Fig. 6 illustrates the joint PDF of $f_{\mathbf{u_jv_j}}(u_j, v_j)$ and the product of the marginal distributions, $f_{\mathbf{u_j}}(u_j)f_{\mathbf{v_j}}(v_j)$, generated by MC simulations. It is qualitatively evident from the figure that there does not exist significant dependence between $\mathbf{u_j}$ and $\mathbf{v_j}$, with simulated mutual information of only 0.003 [34]. Furthermore, Fig. 7 depicts the contour of the error between joint PDF $f_{\mathbf{u_j}}(u_j)f_{\mathbf{v_j}}(v_j)$ and PDF product $f_{\mathbf{u_j}}(u_j)f_{\mathbf{v_j}}(v_j)$ normalized w.r.t. the peak probability on the joint PDF. It is noted that the maximal error is around 7% in a very small region whereas most errors are almost negligible. Moreover, when comparing Figs. 6 and 7, it can be seen that the regions with relatively larger errors have smaller absolute values in Fig. 6, which helps limit the error propagation in (21).

Thus, this independence approximation between u_j and v_j can give us a reasonably accurate estimate of oxide variation

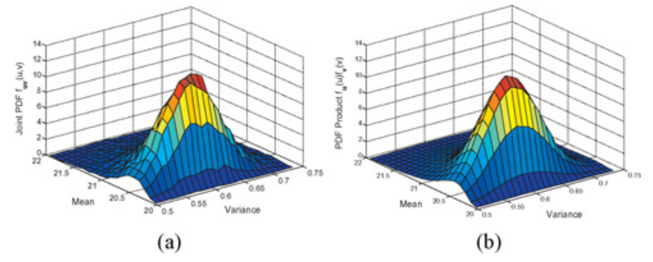

Fig. 6. (a) Joint PDF $f_{\mathbf{u_jv_j}}(u_j, v_j)$. (b) PDF product $f_{\mathbf{u_j}}(u_j)f_{\mathbf{v_j}}(v_j)$.

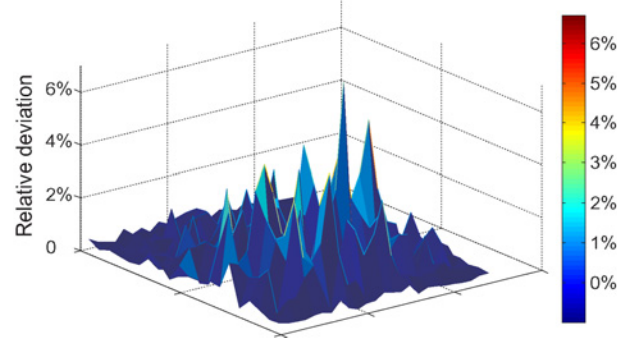

Fig. 7. Contour of the error between joint PDF $f_{\mathbf{u_jv_j}}(u_j, v_j)$ and PDF product $f_{\mathbf{u_j}}(u_j)f_{\mathbf{v_j}}(v_j)$.

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 $f_{\mathbf{u_j}}(u_j)$ and $f_{\mathbf{v_j}}(v_j)$ as follows:

$$\begin{aligned} R_c(t) &= 1 - N \\ &\quad + \sum_{j=1}^{N} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-A_jg(u_j,v_j)} f_{\mathbf{u_j}}(u_j)f_{\mathbf{v_j}}(v_j)du_jdv_j. \end{aligned} \tag{28}$$

Now the BLOD sample mean u_j is a sum of normal random variables, therefore $f_{\mathbf{u_j}}(u_j)$ can be characterized by distribution of a normal random variable and analytically computed. However, BLOD sample variance v_j in (24) is a quadratic expression of normal random variables. Such an expression is commonly found in several multivariate statistics application and is referred to as quadratic normal form [32], [33]. In statistics literature [32], several techniques have been proposed to accurately estimate the distribution function of quadratic normal form. In this paper, we implemented a computationally efficient method given in [33] to estimate the distribution of $f_{\mathbf{v_j}}(v_j)$ using a χ^2 approximation as follows:

$$v_j \sim v_{j,0} + \hat{a}\chi_{\hat{b}}^2 \tag{29}$$

where

$$\begin{aligned} \hat{a} &= \sum_{i=1}^{n}\sum_{k=1}^{n} v_{j,i,k}^2 / \sum_{i=1}^{n} v_{j,i,i} \\ \hat{b} &= (\sum_{i=1}^{n} v_{j,i,i})^2 / \sum_{i=1}^{n}\sum_{k=1}^{n} v_{j,i,k}^2. \end{aligned} \tag{30}$$

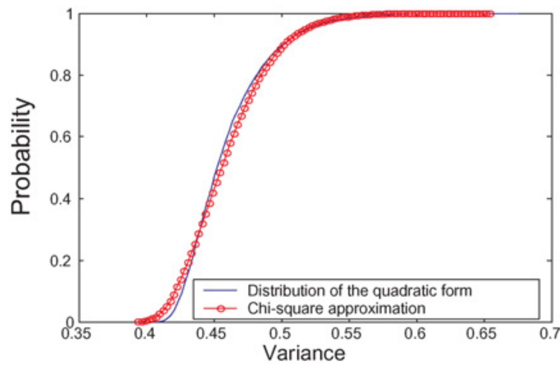

Fig. 8. Curves for the distribution of the quadratic form and its χ^2 approximation.

Procedure: *Full Chip OBD Reliability Analysis*
Input: Number of devices, block-level configuration and related profiles, principal components for (2), inter- and intra-die variation profiles, spatial correlation profile. **Output:** OBD reliability.
1: Characterize u_j and v_j for each BLOD of a given design using (22) and (24); **2:** Divide the integration domain for a double integration to $l_0 \times l_0$ sub-domains; **3:** Compute the sample point pairs (u_j, v_j) for each sub-domain; **4:** **For** each sample point pair **do** **5:** Analytically compute $e^{A_j g(u_j,v_j)}$ using (17); **6:** Evaluate the PDF product $f_{\mathbf{u_j}}(u_j) f_{\mathbf{v_j}}(v_j)$ for the sample point; **7:** **End for** **8:** Compute the overall reliability with the integral sum;

Fig. 9. Algorithm for process variation and temperature-aware full chip OBD reliability analysis.

In Fig. 8, we compare the CDF of the distribution of quadratic normal form of a v_j by MC simulation and its χ^2 approximation. It is apparent that the computationally efficient χ^2 representation is in good agreement with the MC simulation result.

In this manner, the marginal distributions $f_{\mathbf{u_j}}(u_j)$ and $f_{\mathbf{v_j}}(v_j)$ of u_j and v_j can be analytically found for the given process variation model of oxide thickness. Using $f_{\mathbf{u_j}}(u_j)$ and $f_{\mathbf{v_j}}(v_j)$, the overall reliability distribution function can be computed by evaluating N 2-D numerical integrations as in (28).

D. Overall Algorithm

The overall algorithm of the proposed approach is summarized in Fig. 9. Given the principal components as well as the oxide thickness variation profiles, we can characterize u_j and v_j for each BLOD using (22) and (24). Then we divide the integration domain for (28) to $l_0 \times l_0$ sub-domains. Since the joint PDF rapidly decreases to 0 beyond a narrow domain, as illustrated in Fig. 4, $l_0 = 10$ is already a reasonable number for accurate integral sum evaluation, which is further confirmed by the experimental results in Section V. 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.

It is noted that PCA is a pre-processing step. Thus, we do not include it in the complexity analysis as it is performed only once and can be shared with other statistical analysis tools. The overall complexity is $O(N(n^2+l_0^2))$, where N is the number of blocks, n^2 is the number of principal components, and l_0^2 is the number of sub-domains for integration. Unlike the straightforward approach, the computation complexity only depends on the number of temperature-uniform blocks instead of the total number of devices on the chip. Since the number of blocks N is usually much smaller than the total number of devices, it is therefore extremely computationally efficient in comparison to MC analysis, whose complexity heavily depends on the number of devices. Moreover, as temperature and supply voltage are used as the inputs in our model, the correlation of the temperature/voltage profiles between the blocks are naturally captured in the analysis.

E. Fast Computation Using a Hybrid Analytical/Table Look-Up Method

At design time, it is common for designers to repeatedly evaluate the reliability of the same design with different setup and application profiles. Different setup/application profiles may lead to different device-level reliability parameters α and b and hence require computing the integrations again. Although the formulation in (28) significantly reduces the computation complexity, we may achieve further speed up by combining this analytical model with a table look-up method. The pre-calculated look-up table only needs to be computed once for a particular design and can be used for various setup/application profiles or embedded into a dynamic system for reliability monitoring that usually requires very fast response. This could be a possible application of the proposed statistical reliability analysis.

It is usually important to select approbate variables for the look-up table. Equation (28) is comprised of N double integrals. Take the j_{th} integral, e.g., as follows:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-A_j g(u_j,v_j)} f_{\mathbf{u_j}}(u_j) f_{\mathbf{v_j}}(v_j) du_j dv_j \qquad (31)$$

where $g(u_j, v_j) = e^{\ln(\frac{t}{\alpha_j})b_j u_j + (\ln(\frac{t}{\alpha_j}))^2 b_j^2 v_j/2}$. Since u_j and v_j are integration variables, they will be eliminated after the integration is computed. Thus, the result of (31) is determined by A_j and the parameters of t/α_j and b_j in $g(u_j, v_j)$, as shown in (17). Once the chip is designed, A_j appears to be a constant for the j_{th} functional block. Then, with $\ln(t/\alpha)$ and b acting as indices, we can construct a 2-D look-up table to compute the double integral for each block.[5]

Given such a look-up table, the system reliability at any time t under certain temperature/voltage conditions can be computed using bilinear interpolation according to the indices of $\ln(t/\alpha)$ and b. For N functional blocks, we have N look-up tables, with $n_\alpha \times n_b$ entries in each table, where $n_\alpha(=100)$

[5] All the look-up tables for different functional blocks share the same indices of $\ln(t/\alpha_j)$ and b_j. The difference in look-up entries among the blocks is due to the different block area A_j.

TABLE II
EXPERIMENT PARAMETER SETUP FOR THE OBD RELIABILITY ANALYSIS

Quantity	Value
z_0, nominal oxide thickness	2.2 nm
VDD_{nom}, nominal supply voltage	1.2 V
$3\sigma_{tot}/z_0$, total variation w.r.t. the mean [36]	4%
$\sigma^2_{global}/\sigma^2_{tot}$, inter-die variance ratio [37]	50%
$\sigma^2_{spa}/\sigma^2_{tot}$, spatially correlated variance ratio [37]	25%
$\sigma^2_{ind}/\sigma^2_{tot}$, independent variance ratio [37]	25%

and n_b(=100) are the number of indices for parameters $\ln(t/\alpha)$ and b, respectively. Experiments in Section V show that the hybrid method can maintain nearly equivalent accuracy to the analytical approach in Section IV-D.

V. EXPERIMENTAL RESULTS

A simple simulation methodology for estimating the critical defect density required for triggering a dielectric breakdown in an ultrathin oxide was originally developed in [6]. Using this methodology, the defect generation relationships for the technology node and the technology dependent parameters of the oxide reliability function model are then obtained from [7]–[9], [27], which are used in the device-level reliability model (4). In practice, such a model can also be characterized from real OBD distributions measured from test capacitors or discrete devices for the required process and technology.

The proposed approach was implemented in MATLAB and tested on six benchmarks (C1–C6) varying from 50K to 0.84M devices. Designs C1–C5 are synthetic circuits that were automatically generated and design C6 is a alpha processor design with 15 functional modules and approximately 0.84M transistors. We then use HotSpot [10] to achieve the temperature profile of the design with Wattch to estimate the functional block power [35]. In the simulation, we consider the inter-die, spatially correlated intra-die, and random components of variation. According to [36], the $3\sigma/u$ ratio for oxide thickness variation is assumed to be 4% for a nominal value of 2.2 nm, and then split to 50% global variation, 25% spatially correlated variation, and 25% independent variation, as estimated in [37]. As the real measurement data for thickness correlation was unavailable, the covariance matrix for thickness variations used in this paper was derived from an exponential decaying function of the respective distance [38]. The correlation distance of exponential correlation function is normalized w.r.t. the chip dimensions. Table II summarizes the parameters to be used in the analysis framework.

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 MC simulations using the same oxide reliability model and thickness variation model. In the spatial correlation model, the relative correlation distance (ρ_{dist}) w.r.t. the chip size is set to 0.5. In Table III, a comparison of lifetime estimation for 1-fault-per-million parts and 10-faults-per-million parts is shown for six designs. The size of the circuit under test in terms of number of devices is given in the second column. The methods used for comparison include:

1) the proposed statistical approach that uses the marginal PDF product to approximate the joint PDF (*abbrev.* st_fast);
2) the proposed statistical approach that constructs the joint PDF in a numerical way using the MC samples of principal components (*abbrev.* st_MC);
3) the fast hybrid analytical/table look-up approach in Section IV-E (*abbrev.* hybrid);
4) the traditional guard-band method that assumes the minimum oxide thickness and worst-case operating temperature across the chip (*abbrev.* guard) [4], [14], [28];
5) MC simulations (*abbrev.* MC).

The criterion of n-fault-per-million parts is a commonly used term in reliability analysis [39], which is defined as the time when the first n out of a million parts fail. Thus, given that reliability requirement R_{req}, we need to compute the lifetime t_{req} from the integration function in (28), which is as follows:

$$\begin{aligned} R_{req} = \; & 1 - N \\ & + \sum_{j=1}^{N} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-A_j e^{\ln(\frac{t_{req}}{\alpha_j}) b_j u_j + (\ln(\frac{t_{req}}{\alpha_j}))^2 b_j^2 v_j / 2}} \\ & \times f_{\mathbf{u_j}}(u_j) f_{\mathbf{v_j}}(v_j) du_j dv_j. \end{aligned} \tag{32}$$

This could be done by first generating the reliability PDF from (28) at different time stamps and then computing t_{req} from the PDF curve for the given quantile R_{req} by interpolation. For the guard-band method, it is a deterministic process and only needs to solve the following equation to achieve the lifetime estimate [4], [14], [28] as follows:

$$R_{req} = \exp(-A \times (\frac{t_{req}}{\alpha_{worst}})^{b_{worst} x_{min}}) \tag{33}$$

where A is the chip area, α_{worst} and b_{worst} are the parameters corresponding to the worst operating temperature, and x_{min} is the minimum oxide thickness. Thus, t_{req} could be analytically expressed as follows:

$$t_{req} = \alpha_{worst} \times (-\frac{\ln(R_{req})}{A})^{\frac{1}{b_{worst} x_{min}}}. \tag{34}$$

As can be seen from columns 3–5 and 6–8, the proposed methods (st_fast, st_MC, hybrid) are in good agreement with the MC simulation, with errors of around 1% on average. Columns 7–11 compare the runtime for three methods. Unlike MC simulation, both our statistical approaches and hybrid approach are able to analyze the circuit in seconds, with significant speed up. It is also worth noticing that the results show very limited difference between the st_fast method, which uses a marginal PDF product to approximate the joint PDF, and the st_MC method that constructs the joint PDF in a numerical way. The latter has a little runtime overhead to achieve around 0.1% accuracy improvement. The overhead is mainly due to the sample generation and histogram construction. However, since the number of principal components (usually fewer than hundreds) is much smaller than the number of devices, the

TABLE III

ACCURACY AND RUNTIME COMPARISON OF THE TEMPERATURE-AWARE STATISTICAL APPROACH USING MARGINAL PDF (ST_FAST) IN SECTION IV-D, THE STATISTICAL APPROACH THAT CONSTRUCTS THE JOINT PDF NUMERICALLY (ST_MC), THE HYBRID ANALYTICAL/TABLE LOOK-UP APPROACH (HYBRID) IN SECTION IV-E, AND TRADITIONAL GUARD-BAND METHOD (GUARD) WITH MC SIMULATION

ckt.	#Device	Lifetime Estimation Error (%) w.r.t. MC								Runtime (s)/Speed Up w.r.t. MC						
		1/million				10/million				st_fast		st_MC		hybrid		MC
		st_fast	st_MC	hybrid	guard	st_fast	st_MC	hybrid	guard							
C1	50K	0.8	0.8	0.1	42	1.2	1.1	1.8	43	1.5	177	3.2	83	0.02	13 350	267
C2	80K	1.5	1.4	0.7	44	1.3	1.3	0.3	43	1.6	238	3.5	109	0.02	17 490	380
C3	0.1M	2.0	1.8	0.2	56	1.8	1.8	2.3	54	1.9	245	4.1	115	0.02	24 120	470
C4	0.2M	2.2	2.1	0.6	51	1.9	1.9	1.3	51	1.9	363	4.2	167	0.02	35 200	702
C5	0.5M	0.2	0.2	3.4	52	0.1	0.1	1.7	52	1.9	837	4.2	371	0.02	77 850	1557
C6	0.84M	0.6	0.5	1.6	55	0.5	0.5	0.8	54	2.0	1183	4.3	534	0.02	115 330	2307
Average		1.24	1.17	1.11	50	1.13	1.12	1.35	49.5		418		230		47 247	

TABLE IV

ACCURACY COMPARISON BETWEEN THE PROPOSED APPROACH IN SECTION IV-D AND MC SIMULATION FOR DIFFERENT CORRELATION DISTANCE

ckt.	Lifetime Estimation Error w.r.t. MC (%)					
	ρ_{dist} = 0.25		ρ_{dist} = 0.5		ρ_{dist} = 0.75	
	1/million	10/million	1/million	10/million	1/million	10/million
C1	2.31	2.95	0.84	1.18	1.00	1.02
C2	2.26	1.98	1.50	1.28	1.28	1.37
C3	3.35	2.72	2.04	1.77	2.17	1.50
C4	3.77	3.51	2.23	1.90	1.96	1.73
C5	1.62	2.06	0.20	0.12	0.76	0.92
C6	1.70	2.18	0.64	0.54	0.86	0.80

TABLE V

ACCURACY COMPARISON BETWEEN THE PROPOSED APPROACH IN SECTION IV-D AND MC SIMULATION FOR DIFFERENT GRID RESOLUTION FOR DESIGN C2

Grid Size	Lifetime Estimation Error (%) w.r.t. MC Simulation					
	ρ_{dist} = 0.25		ρ_{dist} = 0.5		ρ_{dist} = 0.75	
	1/million	10/million	1/million	10/million	1/million	10/million
10 × 10	3.20	3.17	2.96	3.03	2.87	3.24
20 × 20	2.91	3.08	2.05	1.97	3.01	2.92
25 × 25	2.26	1.98	1.50	1.28	1.28	1.37

st_MC method still demonstrates large runtime improvement. In short, the proposed statistical approach in Section IV-D demonstrates around two to three orders of magnitude speed-up for all the designs, whereas MC simulation scales super-linearly with the number of devices. The hybrid approach in Section IV-E is even faster with three to five orders of magnitude speed-up compared with the MC simulation. Meanwhile, it can maintain a similar accuracy. This is an appealing feature for a real system with increasingly larger designs that may require repeated reliability calculation. Unlike the proposed methods with high accuracy, the results of the traditional guard-band method, as shown in columns 6 and 10, are overly pessimistic, with more than 50% estimation inaccuracy.[6]

To verify the robustness of the proposed approach w.r.t. spatial correlation model, we tested our approach for three different values of correlation distance (ρ_{dist}=0.25, 0.5, 0.75). As can be seen from Table IV, the proposed method can still maintain a good accuracy. We also validate the approach by choosing three different resolutions of grid size for design C2. The numerical results are given in Table V and compared to the MC simulation with a spatial correlation model of 25×25 grids. 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. Moreover, even for the coarsest grid, the analysis result can still maintain a high accuracy for different correlation.

We further compare the overall reliability estimation results in Fig. 10 using MC simulation, the proposed temperature-aware statistical approach in Section IV-D, temperature-unaware approach by using the worst-case temperature across the chip, and conventional guard-band approach that assumes minimum oxide thickness across the chip. Fig. 10 shows the failure rate of design C3 during the selected lifetime period and reliability estimation by different methods. The chip lifetime distribution (blue curve) is achieved by simulating the failure time of 10 000 sample chips of C3 in a MC fashion. One can see that for ten-faults-per-million criterion, the temperature-unaware approach, and conventional guard-band lead to 25.1% and 54.3% errors, whereas our temperature-aware approach can achieve an accuracy of 1.8% error and is very close to the result by MC simulation. This clearly exemplifies the necessity

[6]The runtime result is not included in the table, as the guard-band method only involves the solution of (34), which can be analytically calculated.

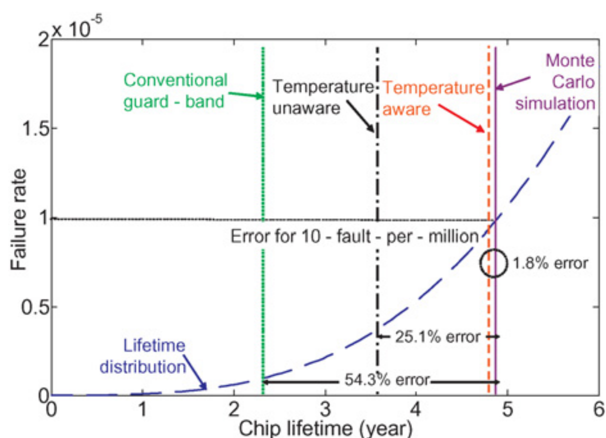

Fig. 10. Errors of the ten-faults-per-million for MC simulation, the proposed temperature-aware approach, temperature-unaware approach using worst-case temperature, and conventional guard-band assuming minimum oxide thickness.

for a process variation and temperature-aware approach for OBD reliability analysis.

VI. Conclusion

This paper proposed a statistical methodology for process and temperature variation-aware chip-level OBD reliability analysis. It is shown that worst-case oxide reliability analysis or temperature-unaware approach may not be adequate to predict chip lifetime accurately. The complexity analysis of the proposed methodology showed that the proposed approach has great scalability to large industrial size circuits. Our simulation results exemplified the accuracy and efficiency of the proposed method.

References

[1] S. Borkar, "Designing reliable systems from unreliable components: The challenges of transistor variability and degradation," *IEEE Micro.*, vol. 25, no. 6, pp. 10–16, Nov.–Dec. 2005.

[2] C. Hu, "Gate oxide scaling limits and projection," in *Proc. IEDM*, 1996, pp. 319–322.

[3] 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*, 2002, pp. 171–174.

[4] 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*, 2006, pp. 18–28.

[5] J. Sune, "New physics-based analytic approach to the thin-oxide breakdown statistics," *IEEE Electron Device Lett.*, vol. 22, no. 6, pp. 296–298, Jun. 2001.

[6] 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*, 1995, pp. 863–866.

[7] E. Wu, D. Harmon, and L. Han, "Interrelationship of voltage and temperature dependence of oxide breakdown for ultrathin oxides," *IEEE Electron Device Lett.*, vol. 21, no. 7, pp. 362–364, Jul. 2000.

[8] E. Wu, J. Sune, W. Lai, E. Nowak, J. McKenna, A. Vayshenkerb, and D. Harmon, "Interplay of voltage and temperature acceleration of oxide breakdown for ultra-thin gate oxides," *Microelectron. Eng.*, vol. 59, nos. 1–4, pp. 25–31, 2001.

[9] R. Degraeve, N. Pangon, B. Kaczer, T. Nigam, G. Groeseneken, and A. Naem, "Temperature acceleration of oxide breakdown and its impact on ultra-thin gate oxide reliability," in *Proc. VLSIT*, 1999, pp. 59–60.

[10] K. Skadron, M. Stan, M. Barcella, A. Dwarka, W. Huang, Y. Li, Y. Ma, A. Naidu, D. Parikh, P. Re, G. Rose, K. Sankaranarayanan, R. Suryanarayan, S. Velusamy, H. Zhang, and Y. Zhang, "Hotspot: Techniques for modeling thermal effects at the processor-architecture level," in *Proc. THERMINICS*, 2002, pp. 169–172.

[11] P. D. Valle and D. Atienza, "Emulation-based transient thermal modeling of 2D/3D systems-on-chip with active cooling," *Microelectron. J.*, vol. 42, no. 4, pp. 564–571, Apr. 2011.

[12] D. Atienza, "Thermal modeling and active cooling for 3D MPSoCs," presented at the Lecture MPSoC Forum, Savannah, GA, Aug. 2009.

[13] J. Friedrich, B. McCredie, N. James, B. Huott, B. Curran, E. Fluhr, G. Mittal, E. Chan, Y. Chan, D. Plass, S. Chu, H. Le, L. Clark, J. Ripley, S. Taylor, J. Dilullo, and M. Lanzerotti, " Design of the Power6 microprocessor," in *Proc. ISSCC*, 2007, pp. 96–97.

[14] E. Karl, D. Sylvester, and D. Blaauw, "Analysis of system-level reliability factors and implications on real-time monitoring methods for oxide breakdown device failures," in *Proc. ISQED*, 2008, pp. 391–395.

[15] K. Skadron, M. Stan, W. Huang, S. Velusamy, K. Sankaranarayanan, and D. Tarjan, "Temperature-aware microarchitecture," in *Proc. ISCA*, 2003, pp. 2–13.

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

[17] C. Zhuo, D. Sylvester, and D. Blaauw, "Process variation and temperature-aware reliability management," in *Proc. DATE*, 2010, pp. 580–585.

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

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

[20] J. Xiong, V. Zolotov, and L. He, "Robust extraction of spatial correlation," in *Proc. ISPD*, 2006, pp. 2–9.

[21] L. Cheng, P. Gupta, C. Spanos, K. Qian, and L. He, "Physically justifiable die-level modeling of spatial variation in view of systematic across wafer variability," in *Proc. DAC*, 2009, pp. 104–109.

[22] K. Qian and C. J. Spanos, "A comprehensive model of process variability for statistical timing optimization," *Proc. SPIE*, vol. 6925, pp. 1–11, 2008.

[23] C. Zhuo, K. Agarwal, D. Sylvester, and D. Blaauw, "Active learning framework for post-silicon variation extraction and test cost reduction," in *Proc. ICCAD*, 2010, pp. 508–515.

[24] A. Agarwal, D. Blaauw, V. Zolotov, S. Sundareswaran, M. Zhou, K. Gala, and R. Panda, "Statistical delay computation considering spatial correlations," in *Proc. ASPDAC*, 2003, pp. 271–276.

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

[26] E. Avni and J. Shappir, "A model for silicon-oxide breakdown under high field and current stress," *J. Appl. Phys.*, vol. 64, pp. 734–742, Jun. 1988.

[27] J. Stathis, "Physical and predictive models of ultra thin oxide reliability in CMOS devices and circuits," *IEEE Trans. Devices Mater. Reliab.*, vol. 1, no. 1, pp. 43–59, Mar. 2001.

[28] J. Sune and E. Y. Wu, "Statistics of successive breakdown events in gate oxides," *IEEE Electron Device Lett.*, vol. 24, no. 4, pp. 272–274, Apr. 2003.

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

[30] J. Fang and S. Sapatnekar, "Scalable methods for the analysis and optimization of gate oxide breakdown," in *Proc. ISQED*, 2010, pp. 638–645.

[31] P. Bhatnagar, S. Dhariwal, and G. Srivastava, "Non-uniformities in ultra-thin oxide layers and their effect on the properties of tunnel MOS, MOM, and SOS devices," *Physica Status Solidi (A)*, vol. 67, no. 1, pp. 305–311, 1981.

[32] J. P. Imhof, "Computing the distribution of quadratic forms in normal variables," *Biometrika*, vol. 48, nos. 3–4, pp. 419–426, 1961.

[33] K.-H. Yuan and P. M. Bentler, "Two simple approximations to the distributions of quadratic forms," *British J. Math. Statist. Psychol.*, vol. 63, no. 2, pp. 273–291, 2010.

[34] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[35] D. Brooks, V. Tiwari, and M. Martonosi, "Wattch: A framework for architectural-level power analysis and optimizations," in *Proc. ISCA*, 2000, pp. 83–94.

[36] International Technology Roadmap for Semiconductors, *Update: Process Integration, Devices, and Structures*, 2008.

[37] S. Reda and S. Nassif, "Analyzing the impact of process variations on parametric measurements: Novel models and applications," in *Proc. DATE*, 2009, pp. 375–380.

[38] F. Liu, "A general framework for spatial correlation modeling in VLSI design," in *Proc. DAC*, 2007, pp. 817–822.

[39] W. Meeker and L. Escobar, *Statistical Methods for Reliability Data*. New York: Wiley, 1998.

Cheng Zhuo (S'06) received the B.S. and M.S. degrees in information science and electronic engineering from Zhejiang University, Hangzhou, China, in 2005 and 2007, respectively, and the Ph.D. degree in computer science and engineering from the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, in 2010.

He was a Summer Intern with Mentor Graphics, San Jose, CA. He is currently working with Intel Corporation, Hillsboro, OR. His current research interests include power integrity issues, timing analysis, design for reliability, and variability-aware optimization.

Kaviraj Chopra received the B.E. degree in instrumentation and control from Gujarat University, Ahmedabad, India, in 2001, the M.S. degree in electrical and computer engineering from the University of Arizona, Tucson, in 2004, and the Ph.D. degree in computer science from the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, in 2008.

He was with the National Optical and Astronomical Observatory, Tucson, AZ, in 2003. He was a Summer Intern with IBM Corporation, Austin, TX, in 2005, and with Synopsys, Mountain View, CA, in 2006. He is currently working as a PnR Engineer with Mentor Graphics, San Jose, CA.

Dennis Sylvester (S'95–M'00–SM'04–F'11) received the Ph.D. degree in electrical engineering from the University of California at Berkeley (UC-Berkeley), Berkeley, where his dissertation was recognized with the David J. Sakrison Memorial Prize as the most outstanding research in the Department of Electrical Engineering and Computer Science, UC-Berkeley.

He is currently a Professor of electrical engineering and computer science with the University of Michigan, Ann Arbor, and the Director of the Michigan Integrated Circuits Laboratory, Ann Arbor, a group of ten faculty and more than 60 graduate students. He was a Research Staff Member with the Advanced Technology Group, Synopsys, Mountain View, CA, and Hewlett-Packard Laboratories, Palo Alto, CA, and a Visiting Professor in electrical and computer engineering at the National University of Singapore, Singapore. He has published over 300 articles along with one book and several book chapters. His current research interests include the design of millimeter-scale computing systems and energy-efficient near-threshold computing for a range of applications. He holds seven U.S. patents. He also is a consultant and technical advisory board member for electronic design automation and semiconductor firms in these areas. He co-founded Ambiq Micro, Ann Arbor, a fabless semiconductor company developing ultralow power mixed-signal solutions for compact wireless devices.

Dr. Sylvester received the NSF CAREER Award, the Beatrice Winner Award at ISSCC, the IBM Faculty Award, the SRC Inventor Recognition Award, and eight best paper awards and nominations. He is the recipient of the ACM SIGDA Outstanding New Faculty Award and the University of Michigan Henry Russel Award for distinguished scholarship. He has served on the technical program committees of major design automation and circuit design conferences, the Executive Committee of the ACM/IEEE Design Automation Conference, and the Steering Committee of the ACM/IEEE International Symposium on Physical Design. He is currently an Associate Editor for the IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN. He was an Associate Editor for the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION SYSTEMS. He is a member of ACM and Eta Kappa Nu.

David Blaauw (M'00–SM'07) received the B.S. degree in physics and computer science from Duke University, Durham, NC, in 1986, and the M.S. and Ph.D. degrees in computer science from the University of Illinois, Urbana, in 1988 and 1991, respectively.

Until August 2001, he was with Motorola, Inc., Austin, TX, as a Manager of the High-Performance Design Technology Group. Since August 2001, he has been on the Faculty of the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, where he is currently a Professor. His current research interests include very large-scale integration design with particular emphasis on ultralow-power and high-performance design.

Dr. Blaauw was the Technical Program Chair and the General Chair for the International Symposium on Low Power Electronics and Design, and the Technical Program Co-Chair and Executive Committee Member of the ACM/IEEE Design Automation Conference. He is currently a Technical Program Committee Member of the International Solid-State Circuits Conference.